CAUSAL DISCOVERY UNDER NON-STATIONARY FEEDBACK

by

Eric V. Strobl

B.A. in Molecular & Cell Biology, University of California at Berkeley, 2009

B.A. in Psychology, University of California at Berkeley, 2009

M.S. in Biomedical Informatics, University of Pittsburgh, 2015

Submitted to the Graduate Faculty of
the Department of Biomedical Informatics in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2017

UNIVERSITY OF PITTSBURGH

BIOMEDICAL INFORMATICS DEPARTMENT

This dissertation was presented

by

Eric V. Strobl

It was defended on

June 23, 2017

and approved by

Shyam Visweswaran, MD, PhD, Associate Professor

Peter L. Spirtes, PhD, Professor

Kun Zhang, PhD, Assistant Professor

Douglas P. Landsittel, PhD, Professor

Gregory F. Cooper, MD, PhD, Professor

Dissertation Director: Shyam Visweswaran, MD, PhD, Associate Professor

CAUSAL DISCOVERY UNDER
NON-STATIONARY FEEDBACK

Eric V. Strobl, PhD

University of Pittsburgh, 2017

Causal discovery algorithms help investigators infer causal relations between random variables using observational data. In this thesis, I relax the acyclicity and stationary distribution assumptions imposed by the Fast Causal Inference (FCI) algorithm, a constraint-based causal discovery method allowing latent common causes and selection bias. I provide two major contributions in doing so. First, I introduce a representation of causal processes called Continuous time Markov processes with Jump points (CMJs) which can model continuous time, feedback loops, and non-stationary distributions. Second, I characterize constraint-based causal discovery under the CMJ framework using a data type which I call *mixture data*, or data created by sampling from a variety of unknown time points from the CMJ. The CMJ may for example correspond to a disease process, and the samples in a mixture dataset to cross-sectional data of patients at different stages in the disease. I finally propose a sound modification of FCI called the Fast Causal Inference with Feedback (F^2CI) algorithm which uses conditional independence testing and conditional mixture modeling to infer causal structure from mixture data even when feedback loops, non-stationary distributions, selection bias and/or latent variables are present. Experiments suggest that the F^2CI algorithm outperforms FCI by a large margin in correctly identifying causal relations when non-stationary distributions and/or feedback loops exist.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to thank my family for supporting me through my education. I would also like to thank all of the members of my PhD committee for helping me improve this thesis.

Research reported in this thesis was supported by grant U54HG008540 awarded by the National Human Genome Research Institute through funds provided by the trans-NIH Big Data to Knowledge initiative. The research was also supported by the National Library of Medicine of the National Institutes of Health under award numbers T15LM007059 and R01LM012095. Finally, funding was also received from the National Institute of General Medical Sciences of the National Institutes of Health under award number T32GM008208. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

1.0 THE PROBLEM

Scientists typically infer causal relations from experimental data. However, they often encounter insurmountable difficulties or unethical scenarios when trying to run experiments. A classic example involves testing the hypothesis of whether smoking tobacco causes lung cancer in human [Spirtes et al., 2000]. Although several observational retrospective and prospective studies beginning in the 1950s reported a strong correlation between smoking and lung cancer, no scientist could ethically design a randomized controlled experiment which forced a random group of people to smoke for years. As a result, many investigators, including the prominent statistician Sir Ronald Fisher, advocated that smoking may not cause lung cancer because other factors could potentially explain the correlation; unmeasured genetic factors could for example cause predispositions to both smoking and lung cancer. The medical community finally reached the conclusion that smoking does in fact cause lung cancer only after painstakingly considering many other variables and performing extensive animal experiments for decades (see Appendix Section A for further details). However, scientists would ideally liked to have discovered this causal relation earlier from the original observational data in order to quickly inform the general populace about the consequences of smoking and therefore save more lives.

Fortunately, many algorithms currently exist for discovering causal relationships from observational data. A large group of these algorithms work by inferring "causal graphs", or a graph where an edge between X_i and X_j indicates a specific type of causal relation between the two variables. Examples of the most well-known causal graph discovery algorithms include PC [Spirtes et al., 2000], FCI [Spirtes et al., 2000, Zhang, 2008] and CCD [Richardson, 1996]. However, these algorithms, as well as all others which I am aware of, either impose assumptions which typically do not apply to biomedical processes or require data types

which medical investigators cannot readily obtain. For example, PC assumes an underlying acyclic causal process (i.e., the causal process does not contain feedback loops) and causal sufficiency (i.e., the data contains all common causes). Most formulations of acyclic causal processes also imply the stationary distribution assumption (i.e., a joint distribution which does not change over time). Note that many biological pathways contain feedback loops and hence acyclic causal graphs often do not apply in medicine without additional time information. Many disease processes also violate the stationary distribution assumption as they evolve over time, and datasets may not contain all common causes. Another algorithm called FCI fortunately eliminates the causal sufficiency assumption and allows selection bias (i.e., variable conditioning due to, for example, inclusion criteria), but it also assumes an underlying acyclic causal graph. The CCD algorithm allows cycles but assumes linear causal relations between the variables and stationary distributions. Thus, PC, FCI and CCD all make assumptions regarding acyclicty and/or stationarity which may not apply to biomedical causal processes.

Given the paramount importance of causal inference in medicine, we require an algorithm which only imposes realistic assumptions and also uses a data type which medical scientists can easily collect. In this thesis, I focus on developing an algorithm which relaxes FCI's acyclicity and stationary distribution assumptions. I also introduce an associated causal framework needed to rigorously justify the algorithm in the following fashion. First, I provide necessary background material in Chapter 3. I then introduce a causal framework called Continuous time Markov processes with Jump points (CMJs) in Chapter 4 which allows feedback loops and non-stationary distributions in continuous time. In Chapter 5, I propose the Fast Casual Inference with Feedback (F^2CI) algorithm which identifies causal structure using mixture data collected from the CMJ without assuming causal sufficiency; here, mixture data refers to data collected at random time points from a CMJ, since I believe that it is unrealistic to assume that scientists can sample from a CMJ at known time points when performing passive observation. Finally, I find that F^2CI outperforms FCI by a large margin in Chapter 6 when non-stationary distributions and/or feedback loops exist. I therefore believe that the algorithm covers a wide variety of realistic scenarios and may serve as a useful tool for causal discovery.

2.0 RELATED WORK

Feedback loops and non-stationary distributions arise in causal processes appearing in nature. For example, biologists have identified feedback loops in gene regulatory networks which induce distributions that change over continuous time [Mitrophanov and Groisman, 2008]. Fortunately, investigators have proposed Markovian-based models which incorporate (a) feedback loops, (b) non-stationary distributions and/or (c) continuous time in order to handle these situations. However, most of the models cannot incorporate all three criteria simultaneously.

Dynamic Bayesian networks (DBNs) for example incorporate feedback loops and non-stationary distributions but do not incorporate continuous time [Dagum et al., 1991, 1992, 1995]. Instead, DBNs model time as discrete time steps. Moreover, methods used to learn DBNs require many samples obtained from the exact time step in the underlying DBN model. Investigators however often have trouble obtaining samples from the exact same time step in practice. Consider for instance a longitudinal dataset containing samples from patients with a particular illness. Here, we cannot ensure that the samples obtained in each wave correspond to the exact same time step in the underlying causal or disease process because some patients may be late in the disease process while others may be early. We therefore would ideally like to model time as a continuous random variable rather than as a discrete deterministic variable.

Investigators have fortunately relaxed the discrete deterministic time assumption in two main ways. One way involves imposing a stationary distribution assumption. This assumption allows investigators to ignore the effect of time, because any independent sample from any time point corresponds to an independent sample from the same distribution. Investigators have utilized this strategy in the context of DBNs as well as more generally in the

context of Structural Equation Models with Independent Errors (SEM-IEs).

SEM-IEs specifically allow feedback loops and continuous time, but the SEM-IE models used in the causality literature require stationary distributions [Spirtes, 1995, Richardson, 1996, Lacerda et al., 2008, Hyttinen et al., 2013]. These stationary distributions in turn carry several limitations. First, stationary distributions arising from non-linear SEM-IEs with feedback loops generally do not satisfy the global directed Markov property, so their utility in causal graph discovery remains uncertain [Spirtes, 1995, Richardson, 1996]. Second, recall that the CCD algorithm assumes that the joint distribution satisfies the global directed Markov property with respect the graph associated with an SEM-IE, so we cannot use the algorithm to learn causal graphs associated with non-linear SEM-IEs in the general case. Third, SEM-IEs associated with feedback loops and stationary distributions suffer from a causal interpretability issue, since the do-operator may no longer have a straightforward interpretation under stationarity [Dash, 2005].

Another line of work has fortunately focused on dropping the stationary distribution assumption by utilizing a time index. However, most authors do not allow feedback loops. In particular, investigators have proposed augmenting constraint-based methods for acyclic causal graph discovery with a time index in order to recover causal structure [Zhang et al., 2017]. Other investigators have suggested capitalizing on non-stationarity to recover acyclic causal structure beyond the Markov equivalence class by utilizing prediction invariance; here, the authors assume structural equations that remain invariant in time but independent errors that may vary with time [Peters et al., 2016, Ghassami et al., 2017]. Both of the aforementioned methods nonetheless require an exact time index which may not be available in many applications (such as in the aforementioned medical dataset).

Now one causal framework known as continuous time Bayesian networks (CTBNs) exists for modeling non-stationary distributions, feedback loops and continuous time simultaneously [Nodelman et al., 2002, 2003]. However, like methods which utilize a time index in order to handle non-stationary distributions, learning CTBNs also requires datasets with time information. Specifically, datasets used to learn CTBNs contain i.i.d. *trajectories* as opposed to i.i.d. random variable values at time points; here, a trajectory corresponds to the evolution of the values of a random variable across time [Nodelman et al., 2003,

2005]. Clearly, obtaining many trajectories may not be possible in practice, since obtaining such samples is much more involved than obtaining values each at singular time points. CTBNs also currently require discrete random variables and model the jump points, or sudden changes of the values of random variables, using a parametric distribution (typically exponential). Many real datasets nonetheless contain continuous random variables and the distribution over jump points may not necessarily follow a parametric model in practice. Finally, algorithms used to learn CTBNs thus far assume causal sufficiency and no selection bias [Nodelman et al., 2003, 2005, Gopalratnam et al., 2005].

In this thesis, I improve upon previous approaches by first introducing a new causal framework which allows non-stationary distributions, feedback loops and continuous time simultaneously just like CTBNs. The framework however does not require a parametric distribution over jump points. Moreover, the framework follows more naturally from acyclic Bayesian networks than CTBNs, since the proposed framework essentially corresponds to an acyclic Bayesian network embedded in continuous time. Second, I propose a corresponding algorithm which does not require datasets composed of trajectories; in fact, the algorithm does not require any time information in order to learn the underlying causal model. The algorithm also does not require causal sufficiency or no selection bias in order to remain sound. I therefore believe that the work described herein represents a more realistic and practical strategy for causal discovery compared to previous approaches.

3.0 BACKGROUND

I now provide the background material required to understand this thesis as follows. In Section 3.1, I first introduce standard graphical terminology used in the causality literature. I then describe the causal interpretation of graphs in Section 3.2. Next, in Sections 3.3, 3.4 and 3.5, I review three causal discovery algorithms called PC, FCI and RFCI which recover causal graphs using conditional independence information. I then review time dependent stochastic processes, Bayesian networks and structural equation models with feedback in Sections 3.6, 3.7 and 3.9, respectively; I will later use the concept of a time dependent stochastic process to derive the CMJ. Subsequently, I review necessary ideas in mixture modeling in Section 3.10. I finally cover two metrics called informedness and markedness in Section 3.11.

3.1 GRAPHICAL DEFINITIONS

A graph $\mathbb{G} = (\boldsymbol{X}, \mathcal{E})$ consists of a set of vertices $\boldsymbol{X} = \{X_1, X_2, \ldots, X_p\}$ and a set of edges $\mathcal{E}$. The edge set $\mathcal{E}$ may contain the following six edge types: $\rightarrow$ (directed), $\leftrightarrow$ (bi-directed), — (undirected), $\circ\!\!\rightarrow$ (partially directed), $\circ\!\!-$ (partially undirected) and $\circ\!\!-\!\!\circ$ (non-directed). Notice that these six edges utilize three types of endpoints including *tails*, *arrowheads* and *circles*. I also use the endpoint "$*$" as a meta-symbol to denote either a tail, arrowhead or circle.

I call a graph containing only directed edges a *directed graph*. On the other hand, a *mixed graph* contains directed, bi-directed and undirected edges. I say that X_i and X_j are *adjacent* in a graph, if they are connected by an edge independent of the edge's type. An

(undirected) path π between X_i and X_j is a set of consecutive edges (also independent of their types) connecting the variables such that no vertex is visited more than once. A *directed path* from X_i to X_j is a set of consecutive directed edges from X_i to X_j in the direction of the arrowheads. A *cycle* occurs when a path exists between X_i and X_j, and X_j and X_i are adjacent. More specifically, a directed path from X_i to X_j forms a *directed cycle* with the directed edge $X_j \to X_i$ and an *almost directed cycle* with the bi-directed edge $X_j \leftrightarrow X_i$.

I say that X_i is an *ancestor* of X_j (and X_j is a *descendant* of X_i), if there exists a directed path from X_i to X_j or $X_i = X_j$. Similarly, X_i is a *parent* of X_j, if there exists a directed edge from X_i to X_j. I say that X_i is a *spouse* of X_j, and X_j is also a spouse of X_i, if there exists a bi-directed edge between X_i and X_j. I denote the set of ancestors, descendants, and parents of X_i as $\boldsymbol{An}(X_i)$, $\boldsymbol{De}(X_i)$ and $\boldsymbol{Pa}(X_i)$, respectively. I also apply these three definitions to a set of vertices $\boldsymbol{W} \subseteq \boldsymbol{X}$ as follows:

$$\begin{aligned}
\boldsymbol{An}(\boldsymbol{W}) &= \{X_i | X_i \in \boldsymbol{An}(X_j) \text{ for some } X_j \in \boldsymbol{W}\}, \\
\boldsymbol{De}(\boldsymbol{W}) &= \{X_i | X_i \in \boldsymbol{De}(X_j) \text{ for some } X_j \in \boldsymbol{W}\}, \\
\boldsymbol{Pa}(\boldsymbol{W}) &= \{X_i | X_i \in \boldsymbol{Pa}(X_j) \text{ for some } X_j \in \boldsymbol{W}\}.
\end{aligned}$$

Three vertices that create a cycle form a *triangle*. On the other hand, three vertices $\{X_i, X_j, X_k\}$ form an *unshielded triple*, if X_i and X_j are adjacent, X_j and X_k are adjacent, but X_i and X_k are not adjacent. I call a nonendpoint vertex X_j on a path π a *collider* on π, if both the edges immediately preceding and succeeding the vertex have an arrowhead at X_j. Likewise, I refer to a nonendpoint vertex X_j on π which is not a collider as a *non-collider*. Finally, an unshielded triple involving $\{X_i, X_j, X_k\}$ is more specifically called a *v-structure*, if X_j is a collider on the subpath $\langle X_i, X_j, X_k \rangle$.

I call a directed graph a *directed acyclic graph* (DAG), if it does not contain directed cycles. Every DAG is a type of *ancestral graph*, or a mixed graph that (1) does not contain directed cycles, (2) does not contain almost directed cycles, and (3) for any undirected edge $X_i - X_j$ in $\mathcal{E}$, X_i and X_j have no parents or spouses [Richardson and Spirtes, 2000].

3.2 PROBABILISTIC AND CAUSAL INTERPRETATION OF GRAPHS

A distribution $\mathbb{P}_{\boldsymbol{X}}$ over $\boldsymbol{X}$ satisfies the *Markov property* if $\mathbb{P}_{\boldsymbol{X}}$ admits a density[1] that "factorizes according to the DAG" as follows:

$$f(\boldsymbol{X}) = \prod_{i=1}^{p} f(X_i | \boldsymbol{Pa}(X_i)). \tag{3.1}$$

We can in turn relate (3.1) to a graphical criterion called d-connection. Specifically, if $\mathbb{G}$ is a directed graph in which $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ are disjoint sets of vertices in $\boldsymbol{X}$, then $\boldsymbol{A}$ and $\boldsymbol{B}$ are *d-connected* given $\boldsymbol{C}$ in the directed graph $\mathbb{G}$ if and only if there exists an undirected path π between some vertex in $\boldsymbol{A}$ and some vertex in $\boldsymbol{B}$ such that every collider on π has a descendant in $\boldsymbol{C}$, and no non-collider on π is in $\boldsymbol{C}$. On the other hand, $\boldsymbol{A}$ and $\boldsymbol{B}$ are *d-separated* given $\boldsymbol{C}$ in $\mathbb{G}$ if and only if $\boldsymbol{A}$ and $\boldsymbol{B}$ are not d-connected given $\boldsymbol{C}$ in $\mathbb{G}$. For shorthand, I will sometimes write $\boldsymbol{A} \perp\!\!\!\perp_d \boldsymbol{B} | \boldsymbol{C}$ and $\boldsymbol{A} \not\perp\!\!\!\perp_d \boldsymbol{B} | \boldsymbol{C}$ when $\boldsymbol{A}$ and $\boldsymbol{B}$ are d-separated or d-connected given $\boldsymbol{C}$, respectively. The conditioning set $\boldsymbol{C}$ is called a *minimal separating set* if and only if $\boldsymbol{A} \perp\!\!\!\perp_d \boldsymbol{B} | \boldsymbol{C}$ but $\boldsymbol{A}$ and $\boldsymbol{B}$ are d-connected given any proper subset of $\boldsymbol{C}$.

Now if we have $\boldsymbol{A} \perp\!\!\!\perp_d \boldsymbol{B} | \boldsymbol{C}$, then $\boldsymbol{A}$ and $\boldsymbol{B}$ are conditionally independent given $\boldsymbol{C}$, denoted as $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} | \boldsymbol{C}$, in any joint density factorizing according to (3.1); I refer to this property as the *global directed Markov property*. I also refer to the converse of the global directed Markov property as *d-separation faithfulness*; that is, if $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} | \boldsymbol{C}$, then we have $\boldsymbol{A} \perp\!\!\!\perp_d \boldsymbol{B} | \boldsymbol{C}$. One can in fact show that the factorization in (3.1) and the global directed Markov property are equivalent, so long as the distribution over $\boldsymbol{X}$ admits a density [Lauritzen et al., 1990].

Now *m-connection* in ancestral graphs is a generalization of d-connection in directed graphs. I say that $\boldsymbol{A}$ and $\boldsymbol{B}$ are m-connected given $\boldsymbol{C}$ in the ancestral graph $\mathbb{G}$ if and only if there exists an undirected path π between some vertex in $\boldsymbol{A}$ and some vertex in $\boldsymbol{B}$ such that every collider on π has a descendant in $\boldsymbol{C}$ and no non-collider on π is in $\boldsymbol{C}$. In turn, $\boldsymbol{A}$ and $\boldsymbol{B}$ are m-separated given $\boldsymbol{C}$ in $\mathbb{G}$ if and only if they are not m-connected given $\boldsymbol{C}$ in $\mathbb{G}$.

[1] I will only consider distributions which admit densities in this thesis.

I write $\boldsymbol{X} = \boldsymbol{O} \cup \boldsymbol{L} \cup \boldsymbol{S}$, when a DAG $\mathbb{G} = (\boldsymbol{X}, \mathcal{E})$ contains non-overlapping sets of observable, latent and selection variables. Here, $\boldsymbol{O}$ denotes the observable variables, $\boldsymbol{L}$ the latent variables, and $\boldsymbol{S}$ the selection variables.

A *maximal ancestral graph* (MAG) is an ancestral graph where every missing edge corresponds to a conditional independence relation. One can transform a DAG $\mathbb{G} = (\boldsymbol{O} \cup \boldsymbol{L} \cup \boldsymbol{S}, \mathcal{E})$ into a MAG $\widetilde{\mathbb{G}} = (\boldsymbol{O}, \widetilde{\mathcal{E}})$ as follows. First, for any pair of vertices $\{O_i, O_j\}$, make them adjacent in $\widetilde{\mathbb{G}}$ if and only if there exists an *inducing path* between O_i and O_j in $\mathbb{G}$. I define an inducing path as follows:

Definition 1. *A path π between O_i and O_j is called an inducing path if and only if every collider on π is an ancestor of $\{O_i, O_j\} \cup \boldsymbol{S}$, and every non-collider on π (except for the endpoints) is in $\boldsymbol{L}$.*

Then, for each adjacency $O_i \ast\!\!-\!\!\ast\, O_j$ in $\widetilde{\mathbb{G}}$, place an arrowhead at O_i if $O_i \notin \boldsymbol{An}(O_j \cup \boldsymbol{S})$ and place a tail if $O_i \in \boldsymbol{An}(O_j \cup \boldsymbol{S})$. The resulting MAG $\widetilde{\mathbb{G}}$ encodes the d-separation and d-connection relations in $\mathbb{G}$ among the observed variables conditional on $\boldsymbol{S}$. That is, O_i and O_j are m-separated by $\boldsymbol{W} \subseteq \boldsymbol{O} \setminus \{O_i, O_j\}$ in $\widetilde{\mathbb{G}}$ if and only if they are d-separated by $\boldsymbol{W} \cup \boldsymbol{S}$ in $\mathbb{G}$ [Spirtes and Richardson, 1996]. The global directed Markov property in turn implies that $O_i \perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S})$ in any distribution with a density factorizing according to $\mathbb{G}$. The MAG of a DAG is therefore a kind of marginal graph that does not contain the latent or selection variables, but does contain information about the ancestral relations between the observable and selection variables in the DAG.

3.3 THE PC ALGORITHM

The PC algorithm considers the following problem: suppose that $\mathbb{P}_{\boldsymbol{X}}$ is d-separation faithful to an unknown DAG $\mathbb{G}$. Then, given oracle information about the conditional independencies between any pair of variables X_i and X_j given any $\boldsymbol{W} \subseteq \boldsymbol{X} \setminus \{X_i, X_j\}$ in $\mathbb{P}_{\boldsymbol{X}}$, reconstruct as much of the underlying DAG as possible. The PC algorithm ultimately accomplishes this goal by reconstructing the DAG up to its *Markov equivalence class*, or the set of DAGs

with the same conditional dependence and independence relations between variables in $\boldsymbol{X}$ [Spirtes et al., 2000].

The PC algorithm represents the Markov equivalence class of DAGs using a *completed partially directed acyclic graph* (CPDAG). A *partially directed acyclic graph* (PDAG) is a graph with both directed and undirected edges. A PDAG is *completed* when the following conditions hold: (1) every directed edge also exists in every DAG belonging to the Markov equivalence class of the DAG, and (2) there exists a DAG with $X_i \rightarrow X_j$ and a DAG with $X_i \leftarrow X_j$ in the Markov equivalence class for every undirected edge $X_i - X_j$. Each edge in the CPDAG also has the following interpretation:

(i) An edge (directed or undirected) is absent between two vertices X_i and X_j if and only if there exists some $\boldsymbol{W} \subseteq \boldsymbol{X} \setminus \{X_i, X_j\}$ such that $X_i \perp\!\!\!\perp X_j | \boldsymbol{W}$.

(ii) If there exists a directed edge from X_i to X_j, then $X_i \in \boldsymbol{Pa}(X_j)$.

The PC algorithm learns the CPDAG through a three step procedure. First, the algorithm initializes a fully connected undirected graph and then determines the presence or absence of each undirected edge using the following fact: under d-separation faithfulness, X_i and X_j are non-adjacent if and only if X_i and X_j are conditionally independent given some subset of $\boldsymbol{Pa}(X_i) \setminus X_j$ or some subset of $\boldsymbol{Pa}(X_j) \setminus X_i$. Note that PC cannot differentiate between the parents and children of a vertex from its neighbors using an undirected graph. Thus, PC tests whether X_i and X_j are conditionally independent given all subsets of $\boldsymbol{Adj}(X_i) \setminus X_j$ and all subsets of $\boldsymbol{Adj}(X_j) \setminus X_i$, where $\boldsymbol{Adj}(X_i)$ denotes the vertices adjacent to X_i in $\mathbb{G}$ (a superset of $\boldsymbol{Pa}(X_i)$), in order to determine the final adjacencies; I refer to this sub-procedure of PC as *skeleton discovery*. The PC algorithm therefore removes the edge between X_i and X_j during skeleton discovery if such a conditional independence is found.

Step 2 of the PC algorithm orients unshielded triples $X_i - X_j - X_k$ to v-structures $X_i \rightarrow X_j \leftarrow X_k$ if X_j is not in the set of variables which rendered X_i and X_k conditionally independent in the skeleton discovery phase of the algorithm. The final step of the PC algorithm involves the repetitive application of three orientation rules to replace as many tails as possible with arrowheads [Meek, 1995].

3.4 THE FCI ALGORITHM

I encourage the reader to compare the aforementioned description of the PC algorithm to the following description of the FCI algorithm. Unlike the PC algorithm, the FCI algorithm considers the following more difficult problem: assume that the distribution of $\boldsymbol{X} = \boldsymbol{O} \cup \boldsymbol{L} \cup \boldsymbol{S}$ is d-separation faithful to an unknown DAG. Then, given oracle information about the conditional independence relations between any pair of observables O_i and O_j given any $\boldsymbol{W} \subseteq \boldsymbol{O} \setminus \{O_i, O_j\}$ as well as $\boldsymbol{S}$, infer as many ancestral relations from the underlying DAG as possible. The FCI algorithm ultimately accomplishes this goal by reconstructing a MAG up to its Markov equivalence class.

The FCI algorithm represents the Markov equivalence class of MAGs, or the set of MAGs with the same conditional dependence and independence relations between variables in $\boldsymbol{O}$ given $\boldsymbol{S}$, using a *completed partial maximal ancestral graph* (CPMAG).[2] A *partial maximal ancestral graph* (PMAG) is a graph with directed, bi-directed, undirected, partially directed, partially undirected and non-directed edges. A PMAG is *completed* when the following conditions hold: (1) every tail and arrowhead also exists in every MAG belonging to the Markov equivalence class of the MAG, and (2) there exists a MAG with a tail and a MAG with an arrowhead in the Markov equivalence class for every circle endpoint. Each edge in the CPMAG also has the following interpretation:

(i) An edge is absent between two vertices O_i and O_j if and only if there exists some $\boldsymbol{W} \subseteq \boldsymbol{O} \setminus \{O_i, O_j\}$ such that $O_i \perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S})$. In other words, an edge is absent if and only if there does not exist an inducing path between O_i and O_j.

(ii) If an edge between O_i and O_j has an arrowhead at O_j, then $O_j \notin \boldsymbol{An}(O_i \cup \boldsymbol{S})$.

(iii) If an edge between O_i and O_j has a tail at O_j, then $O_j \in \boldsymbol{An}(O_i \cup \boldsymbol{S})$.

The FCI algorithm again learns the CPMAG through a three step procedure. The algorithm first performs skeleton discovery by starting with a fully connected nondirected graph, and then the algorithm uses the following fact: under d-separation faithfulness, O_i and O_j are non-adjacent in the CPMAG if and only if O_i and O_j conditionally independent given

[2] The CPMAG is also known as a partial ancestral graph (PAG). However, I will use the term CPMAG in parallel to the use of the term CPDAG.

Data: CI oracle
Result: $\mathbb{G}^M$, sepset, $\mathcal{M}$

Form a complete graph $\mathbb{G}^M$ on $\boldsymbol{O}$ with vertices $\circ\!\!-\!\!\circ$
$l \leftarrow -1$
repeat
  Let $l = l + 1$
  repeat
    forall *vertices in* $\mathbb{G}^M$ **do**
      Compute $\boldsymbol{Adj}(O_i)$
    end
    Select a new ordered pair of vertices (O_i, O_j) that are adjacent in $\mathbb{G}^M$ and satisfy $|\boldsymbol{Adj}(O_i) \setminus O_j| \geq l$
    repeat
      Choose a new set $\boldsymbol{W} \subseteq \boldsymbol{Adj}(O_i) \setminus O_j$ with $|\boldsymbol{W}| = l$
      if $O_i \perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S})$ **then**
        Delete the edge $O_i \circ\!\!-\!\!\circ O_j$ from $\mathbb{G}^M$
        Let $\text{sepset}(O_i, O_j) = \text{sepset}(O_j, O_i) = \boldsymbol{W}$
      end
    until *O_i and O_j are no longer adjacent in $\mathbb{G}^M$ or all $\boldsymbol{W} \subseteq \boldsymbol{Adj}(O_i) \setminus O_j$ with $|\boldsymbol{W}| = l$ have been considered*;
  until *all ordered pairs of adjacent vertices (O_i, O_j) in $\mathbb{G}^M$ with $|\boldsymbol{Adj}(O_i) \setminus O_j| \geq l$ have been considered*;
until *all pairs of adjacent vertices (O_i, O_j) in $\mathbb{G}^M$ satisfy $|\boldsymbol{Adj}(O_i) \setminus O_j| \leq l$*;
Form a list $\mathcal{M}$ of all unshielded triples $\langle O_k, \cdot, O_m \rangle$ (i.e., the middle vertex is left unspecified) in $\mathbb{G}^M$ with $k < m$

Algorithm 1: Obtaining an initial skeleton

Data: $\mathbb{G}^M$, sepset, $\mathcal{M}$
Result: $\mathbb{G}^M$

forall *elements $\langle O_i, O_j, O_k \rangle$ in $\mathcal{M}$* **do**
  if $O_j \notin \text{sepset}(O_i, O_k)$ **then**
    Orient $O_i \ast\!\!-\!\!\circ O_j \circ\!\!-\!\!\ast O_k$ as $O_i \ast\!\!\rightarrow O_j \leftarrow\!\!\ast O_k$ in $\mathbb{G}^M$
  end
end

Algorithm 2: Orienting v-structures

Data: $\mathbb{G}^M$, sepset
Result: $\mathbb{G}^M$, sepset, $\mathcal{M}$

forall *vertices* O_i *in* $\mathbb{G}^M$ **do**
  Compute $\boldsymbol{PDS}(O_i)$
  forall *vertices* $O_j \in \boldsymbol{Adj}(O_i)$ **do**
    Let $l = -1$
    repeat
      Let $l = l + 1$
      repeat
        Choose a (new) set $\boldsymbol{W} \subseteq \boldsymbol{PDS}(O_i) \setminus O_j$ with $|\boldsymbol{W}| = l$
        if $O_i \perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S})$ **then**
          Delete edge $O_i \ast\!\!-\!\!\ast O_j$ in $\mathbb{G}^M$
          Let $\text{sepset}(O_i, O_j) = \text{sepset}(O_j, O_i) = \boldsymbol{W}$
        end
      until *O_i and O_j are no longer adjacent in $\mathbb{G}^M$ or all $\boldsymbol{W} \subseteq \boldsymbol{PDS}(O_i) \setminus O_j$ with $|\boldsymbol{W}| = l$ have been considered*;
    until *O_i and O_j are no longer adjacent in $\mathbb{G}^M$ or $|\boldsymbol{PDS}(O_i) \setminus O_j| < l$*;
  end
end
Reorient all edges in $\mathbb{G}^M$ as $\circ\!\!-\!\!\circ$
Form a list $\mathcal{M}$ of all unshielded triples $\langle O_k, \cdot, O_m \rangle$ in $\mathbb{G}^M$ with $k < m$

Algorithm 3: Obtaining the final skeleton in the FCI algorithm

some subset of $\boldsymbol{DS}(O_i, O_j)$ or some subset of $\boldsymbol{DS}(O_j, O_i)$; here, $O_k \in \boldsymbol{DS}(O_i, O_j)$ if and only if $O_i \neq O_k$, and there exists an undirected path π between O_i and O_k such that every vertex on π is an ancestor of $\{O_i, O_j\} \cup \boldsymbol{S}$ and every non-endpoint vertex is a collider on π. The sets $\boldsymbol{DS}(O_i, O_j)$ and $\boldsymbol{DS}(O_j, O_i)$ thus behave like the parent sets in the DAG. Unfortunately, we cannot compute $\boldsymbol{DS}(O_i, O_j)$ or $\boldsymbol{DS}(O_j, O_i)$ from the conditional independence relations among the observed variables, but Spirtes et al. [2000] identified supersets called *possible d-separating sets* $\boldsymbol{PDS}(O_i)$ and $\boldsymbol{PDS}(O_j)$ s.t. $\boldsymbol{DS}(O_i, O_j) \subseteq \boldsymbol{PDS}(O_i)$ and $\boldsymbol{DS}(O_j, O_i) \subseteq \boldsymbol{PDS}(O_j)$ which we can compute:

Definition 2. *Let $\mathbb{G}$ be a graph with the following edge types: $\circ\!\!-\!\!\circ$, $\circ\!\!\rightarrow$, $\leftrightarrow$. Then, $O_j \in \boldsymbol{PDS}(O_i)$ if and only if there exists a path π between O_i and O_j in $\mathbb{G}$ such that for every subpath $\langle O_m, O_l, O_h \rangle$ of π, O_l is a collider on the subpath in $\mathbb{G}$ or $\langle O_m, O_l, O_h \rangle$ is a triangle in $\mathbb{G}$.*

Note that the definition of $\boldsymbol{PDS}(O_i)$ requires some knowledge about the skeleton and the edge orientations. As a result, the FCI algorithm first creates a completely connected non-directed graph and executes the skeleton discovery procedure summarized in Algorithm 1. FCI then orients the unshielded triple $O_i \circ\!\!-\!\!\circ O_j \circ\!\!-\!\!\circ O_k$ as a v-structure $O_i \circ\!\!\rightarrow O_j \leftarrow\!\!\circ O_k$ using Algorithm 2, if O_i and O_k are rendered conditionally independent given some set not including O_j. The resulting graph contains sufficient information for computing $\boldsymbol{PDS}(O_i)$ for every $O_i \in \boldsymbol{O}$ in Algorithm 3. Thus, the FCI algorithm efficiently computes the skeleton by testing whether O_i and O_j are conditionally independent given all subsets of $\boldsymbol{PDS}(O_i) \backslash O_j$ and all subsets of $\boldsymbol{PDS}(O_j) \setminus O_i$ similar to how PC tests for conditional independence given all subsets of $\boldsymbol{Adj}(O_i) \setminus O_j$ and all subsets of $\boldsymbol{Adj}(O_j) \setminus O_i$. If FCI discovers such a subset which renders O_i and O_j conditionally independent, then the algorithm removes the edge between O_i and O_j.

Step 2 of the FCI algorithm involves the orientation of v-structures again using Algorithm 2 but with the new non-directed skeleton. Subsequently, the algorithm replaces as many circle endpoints with arrowheads and tails in step 3 using ten orientation rules as described in [Zhang, 2008].

3.5 THE RFCI ALGORITHM

The FCI algorithm can take too long to complete when the possible d-separating sets grow large. The RFCI algorithm [Colombo et al., 2012] resolves this problem by recovering a graph where the presence and absence of an edge have the following modified interpretations:

(i) The absence of an edge between two vertices O_i and O_j implies that there exists some $\boldsymbol{W} \subseteq \boldsymbol{O} \setminus \{O_i, O_j\}$ such that $O_i \perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S})$.

(ii) The presence of an edge between two vertices O_i and O_j implies that $O_i \not\perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S})$ for all $\boldsymbol{W} \subseteq \boldsymbol{Adj}(O_i) \setminus O_j$ and for all $\boldsymbol{W} \subseteq \boldsymbol{Adj}(O_j) \setminus O_i$.

We encourage the reader to compare these edge interpretations to the edge interpretations of FCI's CPMAG.

The RFCI algorithm proceeds similarly to FCI but with some modifications. First, RFCI creates a completely connected non-directed graph and initiates the skeleton discovery procedure of FCI using Algorithm 1. However, RFCI then directly starts orienting unshielded triples as v-structures using Algorithm 4 by utilizing the following proposition:

Proposition 1. *[Colombo et al., 2012] Suppose d-separation faithfulness holds. Further assume that (1) O_i and O_j are d-separated given $\boldsymbol{W} \cup \boldsymbol{S}$ with $\boldsymbol{W}$ minimal, and (2) O_i and O_k as well as O_j and O_k are d-connected given $\{\boldsymbol{W} \setminus O_k\} \cup \boldsymbol{S}$. Then $O_k \in \boldsymbol{An}(\{O_i, O_j\} \cup \boldsymbol{S})$ if and only if $O_k \in \boldsymbol{W}$.*

The RFCI algorithm also eliminates additional edges using Algorithm 4 by performing additional conditional independence tests when condition (2) of the above proposition is not satisfied. Next, RFCI applies the orientation rules of FCI, but uses a modified orientation rule 4 due to the following proposition:

Proposition 2. *[Colombo et al., 2012] Suppose d-separation faithfulness holds. Let $\pi_{ik} = \{O_i, \ldots, O_l, O_j, O_k\}$ be a sequence of at least four vertices that satisfy the following: (1) O_i and O_k are conditionally independent given $\boldsymbol{W} \cup \boldsymbol{S}$, (2) any two successive vertices O_h and O_{h+1} on π_{ik} are conditionally dependent given $(\boldsymbol{Y} \setminus \{O_h, O_{h+1}\}) \cup \boldsymbol{S}$ for all $Y \subseteq \boldsymbol{W}$, (3) all vertices O_h between O_i and O_j (not including O_i and O_j) satisfy $O_h \in \boldsymbol{An}(O_k)$ and $O_h \notin \boldsymbol{Anc}(\{O_{h-1}, O_{h+1}\} \cup \boldsymbol{S})$, where O_{h-1} and O_{h+1} denote the vertices adjacent to O_h on*

Data: Initial skeleton $\mathbb{G}^M$, sepset, $\mathcal{M}$
Result: $\mathbb{G}^M$, sepset

Let $\mathcal{L}$ denote an empty list
while $\mathcal{M}$ *is non-empty* **do**
  Choose an unshielded triple $\langle O_i, O_j, O_k \rangle$ from $\mathcal{M}$
  if $O_i \perp\!\!\!\perp O_j | \text{sepset}(O_i, O_k) \cup \boldsymbol{S}$ *and* $O_j \perp\!\!\!\perp O_k | \text{sepset}(O_i, O_k) \cup \boldsymbol{S}$ **then**
    Add $\langle O_i, O_j, O_k \rangle$ to $\mathcal{L}$
  end
  else
    for $r \in \{i, k\}$ **do**
      if $O_r \perp\!\!\!\perp O_j | (\text{sepset}(O_i, O_k) \setminus O_j) \cup \boldsymbol{S}$ **then**
        Find a minimal separating set $\boldsymbol{W} \subseteq \text{sepset}(O_i, O_k)$ for O_r and O_j
        Let $\text{sepset}(O_r, O_j) = \text{sepset}(O_j, O_r) = \boldsymbol{W}$
        Add all triples $\langle O_{\min(r,j)}, \cdot, O_{\max(r,j)} \rangle$ that form a triangle in $\mathbb{G}^M$ into $\mathcal{M}$
        Delete from $\mathcal{M}$ and $\mathcal{L}$ all triples containing $(O_r, O_j) : \langle O_r, O_j, \cdot \rangle$, $\langle O_j, O_r, \cdot \rangle$, $\langle \cdot, O_j, O_r \rangle$ and $\langle \cdot, O_r, O_j \rangle$
        Delete edge $O_r \ast\!\!-\!\!\ast O_j$ in $\mathbb{G}^M$
      end
    end
  end
  Remove $\langle O_i, O_j, O_k \rangle$ from $\mathcal{M}$
end
forall *elements* $\langle O_i, O_j, O_k \rangle$ *of* $\mathcal{L}$ **do**
  if $O_j \notin \text{sepset}(O_i, O_k)$ *and both* $O_i \ast\!\!-\!\!\ast O_j$ *and* $O_j \ast\!\!-\!\!\ast O_k$ *are present in* $\mathbb{G}^M$ **then**
    Orient $O_i \ast\!\!-\!\!\circ O_j \circ\!\!-\!\!\ast O_k$ as $O_i \ast\!\!\rightarrow O_j \leftarrow\!\!\ast O_k$ in $\mathbb{G}^M$
  end
end

Algorithm 4: Orienting v-structures in the RFCI algorithm

π_{ik}. *Then the following hold: (1) if* $O_j \in \boldsymbol{W}$, *then* $O_j \in \boldsymbol{An}(O_k \cup \boldsymbol{S})$ *and* $O_k \in \boldsymbol{An}(O_j \cup \boldsymbol{S})$, *and (2) if* $O_j \notin \boldsymbol{W}$, *then* $O_j \notin \boldsymbol{An}(\{O_l, O_k\} \cup \boldsymbol{S})$ *and* $O_k \notin \boldsymbol{An}(O_j \cup \boldsymbol{S})$.

The RFCI algorithm thus speeds up the FCI algorithm by (1) utilizing smaller sets during skeleton discovery with a relaxed interpretation of the presence and absence of edges, and (2) accordingly modifying its remaining steps in order to remain sound.

3.6 TIME DEPENDENT STOCHASTIC PROCESSES

I define a *time dependent stochastic process* as follows:

Definition 3. *Let* $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$ *denote a probability space, and let the set* Θ *represent time. Suppose further that, for each* $t \in \Theta$, *we have* $p \in \mathbb{N}^+$ *random variables, where we define each random variable* $X_i^t : \Omega_i \to \mathbb{R}$ *on* $(\Omega_i, \mathcal{F}_i, \mathbb{P}_{X_i^t})$. *Then, for each* $t \in \Theta$, *we can consider the random p-vector* $\boldsymbol{X}^t : \prod_{i=1}^p \Omega_i \to \mathbb{R}^p$ *defined on* $(\prod_{i=1}^p \Omega_i, \sigma(\prod_{i=1}^p \mathcal{F}_i), \mathbb{P}_{\boldsymbol{X}^t})$. *The function* $\boldsymbol{X} : \Theta \times \prod_{i=1}^p \Omega_i \to \mathbb{R}^p$ *defined by* $\boldsymbol{X}(t, \prod_{i=1}^p \omega_i) = \boldsymbol{X}^t(\prod_{i=1}^p \omega_i)$ *is called a time dependent stochastic process with indexing set* Θ *and is written as* $\boldsymbol{X}^\Theta = \{\boldsymbol{X}^t | t \in \Theta\}$.

Θ is an arbitrary set, countable or uncountable. I therefore consider the random vector $\boldsymbol{X}$ as a function of two variables on the product space $\Theta \times \prod_{i=1}^p \Omega_i$; this is necessary because I do not want to view the time dependent stochastic process as an arbitrary collection of random variables. Any time point $t \in \Theta$ therefore corresponds to the time in the process as opposed to the clock time, or the time a variable was measured according to a regional time zone. Further observe that I have defined $\boldsymbol{X}^t$ on the same measurable space for all $t \in \Theta$, so I may alternatively refer to $\boldsymbol{X}^t$ as $\boldsymbol{X}$ with probability measure $\mathbb{P}_{\boldsymbol{X}^t}$ (or probability distribution $\mathbb{P}_{\boldsymbol{X}^t}$) without ambiguity.

A *time series variable* refers any member of the set $\{X_i^t | X_i^t \in \boldsymbol{X}^t, t \in \Theta\}$, where t may be unobserved for each member. I however follow the convention in the literature and assume that t is known up to a *relative ordering*; in other words, we must know which random variables are observed *before, during and after* all of the other variables in time. For example, I may observe $\boldsymbol{X}$ at time points $\{0, 10, 20\} \subseteq \Theta$. I may not know however that $\boldsymbol{X}^0$

is observed at time point 0, $\boldsymbol{X}^{10}$ at time point 10, and $\boldsymbol{X}^{20}$ at time point 20. I nonetheless must know that all variables in $\boldsymbol{X}^0$ are observed contemporaneously (at the same time) and before both $\boldsymbol{X}^{10}$ and $\boldsymbol{X}^{20}$. Similarly, I must know that all variables in $\boldsymbol{X}^{10}$ are observed contemporaneously, after $\boldsymbol{X}^0$ and before $\boldsymbol{X}^{20}$. Finally, I must know that all variables in $\boldsymbol{X}^{20}$ are observed contemporaneously and after both $\boldsymbol{X}^0$ and $\boldsymbol{X}^{10}$.

A DAG $\mathbb{G} = (\boldsymbol{V} = \{\boldsymbol{X}^{t_1} \cup \boldsymbol{X}^{t_2} \cup \dots\}, \mathcal{E})$ may represent the relative time ordering between time series variables via directed edges directed contemporaneously or forward in time. That is, if $V_i \to V_j$, then it is understood that $V_i \in \boldsymbol{X}^{t_a}$ and $V_j \in \boldsymbol{X}^{t_b}$ such that $t_a \leq t_b$. We however have two exceptions: *feedback loops* and *self-loops*, where we must have the strict inequality $t_a < t_b$. A feedback loop exists in $\mathbb{G}$ when we have a directed *path* from $X_i^{t_a}$ to $X_i^{t_b}$ where $t_a < t_b$. A self-loop more specifically describes a directed *edge* from $X_i^{t_a}$ to $X_i^{t_b}$ where $t_a < t_b$. I also say that the joint distribution over the time series variables is *stationary* over the set of time points $\mathcal{Q} \subseteq \Theta$ if and only if $\boldsymbol{X}^t \overset{d}{=} \boldsymbol{X}^{t'}, \forall t, t' \in \mathcal{Q}$. Finally, I can consider finite dimensional joint distributions which obey the Markov property according to $\mathbb{G}$ and therefore also the global directed Markov property just like with the original DAG.

3.7 BAYESIAN NETWORKS & DYNAMIC BAYESIAN NETWORKS

Rather than describe a DAG $\mathbb{G}$ and its associated factorizable distribution separately, I use the term *Bayesian network* (BN) in order to directly refer to the double $(\mathbb{G} = (\boldsymbol{X}, \mathcal{E}), \mathbb{P}_{\boldsymbol{X}})$, where $\mathbb{G}$ is a DAG, and $\mathbb{P}_{\boldsymbol{X}}$ is a distribution over $\boldsymbol{X}$ with a density that factorizes according to $\mathbb{G}$. On the other hand, a *dynamic Bayesian network* (DBN), on the other hand, refers to the double $(\mathbb{G} = (\boldsymbol{V} = \cup_{i=1}^{q} \boldsymbol{X}^{t_i}, \mathcal{E}), \mathbb{P}_{\boldsymbol{V}})$, where $\Theta = \{t_1, \dots, t_q\}$ and $q \in \mathbb{N}^+$. Moreover, $\mathbb{P}_{\boldsymbol{V}}$ denotes a distribution over $\cup_{i=1}^{q} \boldsymbol{X}^{t_i}$ also with a density that factorizes according to $\mathbb{G}$ [Dagum et al., 1991, 1992, 1995].

3.8 CONTINUOUS TIME BAYESIAN NETWORKS

Note that the term *Continuous Time Bayesian Network* (CTBN) does not correspond to a straightforward extension of the DBN; we do not just consider the double $(\mathbb{G} = (\boldsymbol{V} = \{\boldsymbol{X}^t : t \in \Theta\}, \mathcal{E}), \mathbb{P}_{\boldsymbol{V}})$, where we model time as continuous by setting Θ to some finite interval. Instead, investigators have defined the CTBN using a different approach described in [Nodelman et al., 2002]. Here, we assume that every random variable in $\boldsymbol{V}$ is discrete. We also require:

1. An initial distribution $\mathbb{P}_{\boldsymbol{X}^0}$ whose density factorizes according to a DAG $\mathbb{G}_0$;
2. A time transition probability distribution typically set to the exponential or Erlang-Coxian distribution [Gopalratnam et al., 2005];
3. A second graph $\mathbb{G}_1$ (possibly cyclic) as well as conditional transition matrices $\mathcal{Q}_{\boldsymbol{X}|\boldsymbol{Pa}(X_i)}$, for each variable $X_i \in \boldsymbol{X}$.

The data generating process of CTBN model operates by first sampling according to $\mathbb{P}_{\boldsymbol{X}^0}$, then sampling the transition time t, and finally sampling $\boldsymbol{X}^t$ according to the conditional transition matrices.

3.9 STRUCTURAL EQUATION MODELS WITH FEEDBACK

Consider the double $(\mathbb{G} = (\boldsymbol{X}, \mathcal{E}), \mathbb{P}_{\boldsymbol{X}})$, where $\mathbb{G}$ is a directed graph that may contain cycles. In this case, $\mathbb{P}_{\boldsymbol{X}}$ may not obey the global directed Markov property. We can however impose certain assumptions on $\mathbb{P}_{\boldsymbol{X}}$ such that it does obey the property.

Spirtes [1995] proposed the following assumptions on $\mathbb{P}_{\boldsymbol{X}}$. We say that a distribution $\mathbb{P}_{\boldsymbol{X}}$ obeys a *structural equation model with independent errors* (SEM-IE) with respect to $\mathbb{G}$ if we may describe $\boldsymbol{X}$ as $X_i = g_i(\boldsymbol{Pa}(X_i), \varepsilon_i)$ for all $X_i \in \boldsymbol{X}$ such that X_i is $\sigma(\boldsymbol{Pa}(X_i), \varepsilon_i)$ measurable [Evans, 2016] and $\varepsilon_i \in \boldsymbol{\varepsilon}$. Here, we have a set of jointly independent errors $\boldsymbol{\varepsilon}$, and $\sigma(Y)$ refers to the sigma-algebra generated by the random variable Y.

I provide an example of an SEM-IE below:

$$\begin{aligned} X_1 &= \varepsilon_1, \\ X_2 &= B_{12}X_1 + B_{32}X_3 + \varepsilon_2, \\ X_3 &= B_{43}X_4 + B_{32}X_2 + \varepsilon_3, \\ X_4 &= \varepsilon_4, \end{aligned} \tag{3.2}$$

where $\boldsymbol{\varepsilon}$ denotes a set of jointly independent standard Gaussian error terms, and B is a 4 by 4 coefficient matrix with a diagonal of zeros, since we will not consider self-loops. Notice that the structural equations in (3.2) are linear structural equations, but we can also consider non-linear structural equations.

We can simulate data from an SEM-IE using the *fixed point method*. The fixed point method involves two steps per sample. We first sample the error terms according to their independent distributions and initialize $\boldsymbol{X}$ to some values. Next, we apply the structural equations iteratively until the values of the random variables converge to values which satisfy the structural equations; in other words, the values converge to a fixed point. Note that the values of the random variables may not necessarily converge to a fixed point all of the time for every set of structural equations and error distributions, but I will only consider those structural equations and error distributions which do satisfy this property. Of course, the method terminates all of the time in the acyclic case, since we only need to perform one iteration over the structural equations per sample.

We can perform the fixed point method more efficiently in the linear case by first representing the structural equations in matrix format: $\boldsymbol{X} = B^T\boldsymbol{X} + \boldsymbol{\varepsilon}$. Then, after drawing the values of $\boldsymbol{\varepsilon}$, we can obtain the values of $\boldsymbol{X}$ by solving the following system of equations: $\boldsymbol{X} = (\mathbb{I} - B^T)^{-1}\boldsymbol{\varepsilon}$, where $\mathbb{I}$ denotes the identity matrix.

Spirtes [1995] proved the following regarding *linear SEM-IEs*, or SEM-IEs with linear structural equations:

Theorem 1. *The probability distribution $\mathbb{P}_{\boldsymbol{X}}$ of a linear SEM-IE satisfies the global directed Markov property with respect to the SEM-IE's directed graph $\mathbb{G}$ (acyclic or cyclic).*

The above theorem provided a basis from which Richardson started constructing the Cyclic Causal Discovery (CCD) algorithm [Richardson, 1996] for causal discovery with feedback.

3.9.1 Chain Graphs

Lauritzen and Richardson also introduce the notion of a chain graph which model stationary distributions of causal models with feedback in a similar manner to SEM-IEs [Lauritzen and Richardson, 2002]. A chain graph $\mathbb{G}$ corresponds to a graph with both undirected and directed edges. A distribution associated with a chain graph factorizes as follows:

$$f(\boldsymbol{X}) = \prod_{\xi \in \Xi} f(\boldsymbol{X}_\xi | \boldsymbol{Pa}(\boldsymbol{X}_\xi)), \tag{3.3}$$

where the set Ξ contains the *chain components* of $\mathbb{G}$, or the connected components of an undirected graph obtained by removing all directed edges from $\mathbb{G}$. Note that the chain components of a DAG are all singletons, since the DAG does not contain any undirected edges.

Algorithm 5 represents one possible way of sampling from a chain graph, where I have reused the notation of an SEM-IE in line 6. Notice that the algorithm takes as input an initial instantiation $\boldsymbol{x}_0$ and an ordered set of chain components Ξ. The algorithm then outputs the final sample $\boldsymbol{x}$. Further observe that Algorithm 5 contains an outer and an inner loop. The outer loop cycles over the chain components. On the other hand, the inner loop updates the variables within a chain component ξ until the variables in ξ converge to a fixed point. The sampling procedure of a chain graph therefore bears close resemblance to the sampling procedure of an SEM-IE.

3.10 MIXTURE MODELS

Consider a family of densities $\{f(\boldsymbol{X}|\psi) : \psi \in \Psi\}$ with respect to a measure μ. I call the following density a *mixture density* with respect to the *mixing distribution* $F(\psi)$:

$$f(\boldsymbol{X}) = \int_\Psi f(\boldsymbol{X}|\psi)\ dF(\psi). \tag{3.4}$$

In this thesis, I will focus on *finite mixture densities* which take the following form:

$$f_{\boldsymbol{\theta}}(\boldsymbol{X}) = \sum_{j=1}^{m} \lambda_j f(\boldsymbol{X}|\psi_j) = \sum_{j=1}^{m} \lambda_j f_j(\boldsymbol{X}), \tag{3.5}$$

Data: $\boldsymbol{x}_0$, ordered Ξ

Result: $\boldsymbol{x}$

$\boldsymbol{x} \leftarrow \boldsymbol{x}_0$
forall $\xi \in \Xi$ **do**
  $j \leftarrow 0$
  repeat
    $j \leftarrow j + \text{mod}(\xi)$
    $\boldsymbol{x}_\xi \rightarrow g_\xi(\boldsymbol{Pa}(\boldsymbol{X}_\xi), \varepsilon_\xi)$
  until $\boldsymbol{x}_\xi$ *converges to a fixed point*;
end

Algorithm 5: Sampling a Distribution Associated with a Chain Graph

where $0 < \lambda_j \leq 1$, $\sum_{j=1}^{m} \lambda_j = 1$, $\boldsymbol{\theta} = \{\boldsymbol{\lambda} \cup \boldsymbol{f}\} = \{\lambda_1, \ldots, \lambda_m, f_1, \ldots, f_m\}$, and m denotes the total number of unique densities [Everitt and Hand, 1981]. Each $f_j(\boldsymbol{X})$ may for example correspond to a Gaussian density.

Given a random sample $\boldsymbol{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ from the density (3.5), we can write the likelihood as:

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), \tag{3.6}$$

Investigators usually estimate each mixing proportion λ_j and the parameters of each density $f_j(\boldsymbol{X})$ by maximizing the log-likelihood using the expectation-maximization (EM) algorithm. In fact, a number of authors provide weak conditions that ensure the existence, consistency and asymptotic normality of the maximum likelihood parameter estimates for finite mixtures of densities from an exponential family [Redner and Walker, 1984, Atienza et al., 2007].

3.10.1 The EM Algorithm for Finite Mixture Models

I now describe the EM algorithm for estimating a finite mixture model [Dempster et al., 1977, Benaglia et al., 2009]. Suppose that we can divide the *complete data* $\boldsymbol{c} = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n\}$ into *observed data* and *missing data*; in this situation, we have the complete data vector

$\boldsymbol{c}_i = (\boldsymbol{x}_i, \boldsymbol{z}_i)$ where $\boldsymbol{x}_i$ denotes the observed data and $\boldsymbol{z}_i$ the missing data. Associate $\boldsymbol{x}$ with the log-likelihood $L_{\boldsymbol{x}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\, f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$ and the complete data with the likelihood $h_{\boldsymbol{\theta}}(\boldsymbol{c}) = \prod_{i=1}^{n} h_{\boldsymbol{\theta}}(\boldsymbol{c_i})$ and the log-likelihood $\log\, h_{\boldsymbol{\theta}}(\boldsymbol{c}) = \sum_{i=1}^{n} \log\, h_{\boldsymbol{\theta}}(\boldsymbol{c_i})$.

In Equation (3.5), we can consider the random vector $\boldsymbol{C}_i = (\boldsymbol{X}_i, \boldsymbol{Z}_i)$ where $\boldsymbol{Z}_i = (Z_{ij}, j = 1, \ldots, m)$, and we can view $Z_{ij} \in \{0, 1\}$ as a Bernoulli random variable indicating that individual i comes from component j. Notice however that the constraint $\sum_{j=1}^{m} Z_{ij} = 1$ must hold, since each sample comes from one component. We also have:

$$\mathbb{P}(Z_{ij} = 1) = \lambda_j, \quad (\boldsymbol{X}_i | Z_{ij} = 1) \sim f_j, \quad j = 1, \ldots, m. \tag{3.7}$$

We can therefore write the complete data density evaluated at one sample as:

$$h_{\boldsymbol{\theta}}(\boldsymbol{c_i}) = h_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{z}_i) = \sum_{j=1}^{m} \mathbb{I}_{z_{ij}} \lambda_j f_j(\boldsymbol{x}_i). \tag{3.8}$$

The EM algorithm does not maximize the log-likelihood over the observed data but instead maximizes the following quantity:

$$Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(t)}) = \mathbb{E}\big[\log\, h_{\boldsymbol{\theta}}(\boldsymbol{C}) | \boldsymbol{x}, \boldsymbol{\theta}^{(t)}\big]. \tag{3.9}$$

Here, we take the expectation with respect to the density $k_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{c}|\boldsymbol{x}) = \prod_{i=1}^{n} k_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{z}_i|\boldsymbol{x}_i)$, where $\boldsymbol{\theta}^{(t)}$ denotes the parameters at iteration t. The procedure to transition from $\boldsymbol{\theta}^{(t)}$ to $\boldsymbol{\theta}^{(t+1)}$ takes the following form:

1. Expectation step (E-step): compute $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$,
2. Maximization step (M-step): set $\boldsymbol{\theta}^{(t+1)}$ to $\arg\max_{\boldsymbol{\theta} \in \Theta} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$.

Let us take a closer look at the case of a finite Gaussian mixture. We have the following E and M-steps:

1. E-step: First note that we can compute the following probability conditional on the observed data and $\boldsymbol{\theta}^{(t)}$ by Bayes' theorem:

$$\begin{aligned}\phi_{ij}^{(t)} &\stackrel{\text{def}}{=} \mathbb{P}(Z_{ij}=1|\boldsymbol{x}_i,\boldsymbol{\theta}^{(t)}) \\ &= \frac{\lambda_j^{(t)}\zeta_j^{(t)}(\boldsymbol{x}_i)}{\sum_{j'=1}^m \lambda_{j'}^{(t)}\zeta_{j'}^{(t)}(\boldsymbol{x}_i)}\end{aligned} \tag{3.10}$$

for all $i=1,\ldots,n$ and $j=1,\ldots,m$. Here, ζ_j denotes the normal density with mean and covariance (μ_j,Σ_j). We can now write $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ compactly as follows:

$$\begin{aligned}Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \mathbb{E}\big[\log\, h_{\boldsymbol{\theta}}(\boldsymbol{C})|\boldsymbol{x},\boldsymbol{\theta}^{(t)}\big] = \mathbb{E}\big[\sum_{i=1}^n \log\, h_{\boldsymbol{\theta}}(\boldsymbol{c}_i)|\boldsymbol{x},\boldsymbol{\theta}^{(t)}\big] \\ &= \sum_{i=1}^n \mathbb{E}\big[\log\, h_{\boldsymbol{\theta}}(\boldsymbol{c}_i)|\boldsymbol{x},\boldsymbol{\theta}^{(t)}\big] \\ &= \sum_{i=1}^n\sum_{j=1}^m \mathbb{P}(Z_{ij}=1|\boldsymbol{x}_i,\boldsymbol{\theta}^{(t)})\log\, \lambda_j\zeta_j(\boldsymbol{c}_i) \\ &= \sum_{i=1}^n\sum_{j=1}^m \phi_{ij}^{(t)}\log\, \lambda_j\zeta_j(\boldsymbol{c}_i).\end{aligned} \tag{3.11}$$

2. M-step: We need to perform the following maximization:

$$\begin{aligned}\boldsymbol{\theta}^{(t+1)} &= \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \\ &= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^n\sum_{j=1}^m \phi_{ij}^{(t)}\log\, \lambda_j\zeta_j(\boldsymbol{c}_i).\end{aligned} \tag{3.12}$$

Let us first consider the parameters $\boldsymbol{\lambda}$. We can write the value of each λ_j which maximizes (3.12) in closed from in a format similar to the MLE of the binomial distribution:

$$\lambda_j^{(t+1)} = \frac{\sum_{i=1}^n \phi_{ij}^{(t)}}{\sum_{i=1}^n\sum_{j'=1}^m \phi_{ij'}^{(t)}} = \frac{1}{n}\sum_{i=1}^n \phi_{ij}^{(t)}. \tag{3.13}$$

Likewise, the means and covariances which maximize (3.12) have the following closed form similar to the MLEs of the weighted Gaussian:

$$\begin{aligned}\mu_j^{(t+1)} &= \frac{\sum_{i=1}^n \phi_{ij}^{(t)}\boldsymbol{x}_i}{\sum_{i=1}^n \phi_{ij}^{(t)}}, \\ \Sigma_j^{(t+1)} &= \frac{\sum_{i=1}^n \phi_{ij}^{(t)}(\boldsymbol{x}_i-\mu_j^{(t+1)})(\boldsymbol{x}_i-\mu_j^{(t+1)})^T}{\sum_{i=1}^n \phi_{ij}^{(t)}}.\end{aligned} \tag{3.14}$$

Now the above closed form expressions imply that we can compute both the E and M-steps of the EM algorithm for finite Gaussian mixtures in a short period of time. However, notice that the optimization problem in Equation 3.9 is non-convex in general, so we cannot guarantee that the EM algorithm will always converge to the global maximum of the likelihood; this drawback is unfortunate, but it is shared by many popular clustering algorithms such as k-means.

The EM algorithm carries two desirable properties on the other hand. First, the algorithm increases the complete data log-likelihood after each M-step under weak conditions (i.e., $\log h_{\boldsymbol{\theta}^{(t+1)}}(\boldsymbol{C}) \geq \log h_{\boldsymbol{\theta}^{(t)}}(\boldsymbol{C})$), so the log-likelihood converges monotonically to a local maximum [Dempster et al., 1977, Wu, 1983, Boyles, 1983, Redner and Walker, 1984]. Second, many investigator have noted that the EM algorithm for finite Gaussian mixtures performs very well in practice across a variety of synthetic and real data problems (e.g., [Ortiz and Kaelbling, 1999, Yusoff et al., 2009]). We also indirectly replicate these strong empirical results in our experiments.

3.10.2 Finite Mixtures of Single Response Linear Regressions

We will be particularly interested in estimating the following finite conditional mixture density:

$$f(Y|\boldsymbol{X}) = \sum_{j=1}^{m} \lambda_j f(Y|\boldsymbol{X}, \psi_j) \tag{3.15}$$

Suppose now that we can describe the functional relationship between Y and $\boldsymbol{X}$ in the j^{th} component using the following linear model:

$$Y = \boldsymbol{X}^T \beta_j + \varepsilon_j, \tag{3.16}$$

where $\varepsilon_j \sim \mathcal{N}(0, \sigma_j^2)$. Then, the conditional mixture density admits the following form:

$$f(Y|\boldsymbol{X}) = \sum_{j=1}^{m} \lambda_j \mathcal{N}(Y|\boldsymbol{X}^T \beta_j, \sigma_j^2). \tag{3.17}$$

where $\mathcal{N}(Y|\boldsymbol{X}^T \beta_j, \sigma_j^2)$ denotes the normal density with mean $\boldsymbol{X}^T \beta_j$ and covariance σ_j^2 [Quandt and Ramsey, 1978].

We can again use the EM algorithm to find a local maximum of the expected likelihood [De Veaux, 1989]. The E-step proceeds similarly with the update rule for the finite Gaussian mixture model, but we replace $\zeta_j^{(t)}(\boldsymbol{x}_i)$ in (3.10) with $\zeta(y_i|\boldsymbol{x}_i^T\beta_j^{(t)}, \sigma_j^{2,(t)})$:

$$\phi_{ij}^{(t)} = \frac{\lambda_j^{(t)}\zeta(y_i|\boldsymbol{x}_i^T\beta_j^{(t)}, \sigma_j^{2,(t)})}{\sum_{j'=1}^{m}\lambda_{j'}^{(t)}\zeta(y_i|\boldsymbol{x}_i^T\beta_{j'}^{(t)}, \sigma_{j'}^{2,(t)})}. \tag{3.18}$$

The M-step also proceeds in an analogous fashion. Let $W_j^{(t)} = \text{diag}(\phi_{1j}^{(t)}, \ldots, \phi_{nj}^{(t)})$. Place the i.i.d. samples of $\boldsymbol{X}$ and Y in the rows of the matrix $\underline{\boldsymbol{X}}$ and in the rows of the column vector $\underline{Y}$, respectively. The M-step updates of the β and σ parameters are given by:

$$\begin{aligned}
\beta_j^{(t+1)} &= (\underline{\boldsymbol{X}}^T W_j^{(t)} \underline{\boldsymbol{X}})^{-1} \underline{\boldsymbol{X}}^T W_j^{(t)} \underline{Y}, \\
\sigma_j^{2(t+1)} &= \frac{\left\| \sqrt{W_j^{(t)}}(\underline{Y} - \underline{\boldsymbol{X}}\beta_j^{(t+1)}) \right\|^2}{\text{tr}(W_j^{(t)})},
\end{aligned} \tag{3.19}$$

where $\text{tr}(A)$ means the trace of A and $\|A\|^2 = A^T A$. Notice that the first equation in (3.19) is a weighted least squares (WLS) estimate of β_j, and the second equation resembles the variance estimate used in WLS.

3.11 INFORMEDNESS AND MARKEDNESS

Many investigators analyze algorithmic results using recall and precision:

$$\begin{aligned}
\text{Recall} &= \frac{TP}{TP + FN}, \\
\text{Precision} &= \frac{TP}{TP + FP},
\end{aligned} \tag{3.20}$$

where TP, FP, TN, and FN denote the number of true positives, false positives, true negatives and false negatives, respectively. Notice that we may obtain high recall and high precision by predicting the positive class accurately but guessing the negative class, if the number of positives far outweighs the number of negatives. This deficiency arises for two reasons. First, recall quantifies the proportion of correct predicted positives, while precision quantifies the proportion of correct real positives. Both measures thus ignore performance in correctly

handling negative examples; in fact, both measures do not consider the number of true negatives. Second, precision and recall vary depending on the algorithm bias (proportion of predicted positives; $\frac{TP+FP}{TP+FP+TN+FN}$) and the population prevalence (proportion of real positives; $\frac{TP+FN}{TP+FP+TN+FN}$). Consider for instance a biased algorithm that always guesses positive. If we run the algorithm on a population that has a high prevalence, then we will have both high precision and high recall because FN and FP will both be small. Recall and precision therefore fail to take into account chance level performance particularly when class imbalances exist.

Powers proposed the informedness and markedness measures to correct for the aforementioned shortcomings of precision and recall [Powers]. Define inverse recall and inverse precision similar to recall and precision respectively, but with the positive class and the negative class reversed:

$$\begin{aligned} \text{Inverse Recall} &= \frac{TN}{TN+FP}, \\ \text{Inverse Precision} &= \frac{TN}{TN+FN}. \end{aligned} \tag{3.21}$$

Next, define informedness as a balanced measure of recall and inverse recall, and markedness as a balanced measure of precision and inverse precision as follows:

$$\begin{aligned} \text{Informedness} &= \text{Recall} + \text{Inverse Recall} - 1, \\ \text{Markedness} &= \text{Precision} + \text{Inverse Precision} - 1. \end{aligned} \tag{3.22}$$

Intuitively then informedness and markedness take the negative class more into account than recall and precision by considering the inverse measures.

More deeply, the above two equations have interesting connections to bias and prevalence. We start the argument by defining *chance level* as random guessing at the level of algorithm bias; thus an algorithm guesses positive 90% of the time regardless of the population, if the bias is 90%. We can in turn use bias and prevalence to compute the expected true positive rate (ETPR) at chance level. By independence of the prevalence and bias due to random guessing, we have the product ETPR = Bias $\times$ Prevalence. For example, if the bias is 90% and the prevalence 95%, then we expect an 85.5% true positive rate from random guessing.

Powers showed that we can re-write the equations in 3.22 as follows:

$$\begin{aligned} \text{Informedness} &= \frac{\text{TPR} - \text{ETPR}}{\text{Prevalence} \times (1 - \text{Prevalence})}, \\ \text{Markedness} &= \frac{\text{TPR} - \text{ETPR}}{\text{Bias} \times (1 - \text{Bias})}, \end{aligned} \tag{3.23}$$

where TPR denotes the true positive rate. Notice then that informedness and markedness both yield an expected value of zero when the algorithm performs at chance level regardless of the bias or prevalence level because $\mathbb{E}(TPR) = ETPR$ in this case.

Recall and precision nonetheless do not in general yield an expected value of zero when the algorithm performs at chance level. The expected value for recall and precision at chance level instead varies depending on the underlying bias and prevalence levels. Powers proved this fact by writing the relation between informedness and recall as well as markedness and precision as follows:

$$\begin{aligned} \text{Informedness} &= \frac{\text{Recall} - \text{Bias}}{1 - \text{Prevalence}}, \\ \text{Markedness} &= \frac{\text{Precision} - \text{Prevalence}}{1 - \text{Bias}}. \end{aligned} \tag{3.24}$$

We can thus view informedness and markedness as recall and precision, respectively, after controlling for bias and prevalence. We will find informedness and markedness useful for analyzing the experimental results, since the measures more accurately quantify the degree to which the algorithms are performing better than chance level compared to recall and precision.

Finally note that many investigators often maximize the Area Under the Curve (AUC) in the machine learning, epidemiology and statistics literatures. We can in fact write the following:

$$\begin{aligned} \text{Informedness} &= \text{Recall} + \text{Inverse Recall} - 1 \\ &= \text{Sensitivity} + \text{Specificity} - 1, \\ \text{AUC} &= \frac{\text{Sensitivity} + \text{Specificity}}{2} \\ &= \frac{\text{Informedness} + 1}{2}. \end{aligned} \tag{3.25}$$

Notice here that the AUC corresponds to the area of a trapezoid, which is equivalent to the definition of the AUC when it is dictated by a single point. We thus conclude that maximizing the single point AUC is equivalent to maximizing informedness. However, observe that the

AUC does not take precision into account, so some investigators prefer to maximize the Area Under the Precision-Recall Curve (AUPRC) instead by essentially replacing specificity with precision. We may specifically write the single point Area Under the Precision and 1-Recall Curve (AUPR^{1}C) as follows:

$$\begin{aligned} \text{AUPR}^1\text{C} &= \frac{\text{Recall} + \text{Precision}}{2} \\ &= \frac{\text{Sensitivity} + \text{Precision}}{2}, \end{aligned} \tag{3.26}$$

but this again does not correct for bias and prevalence. The same conclusion holds for single point AUPRC with equations slightly less analogous to those of single point AUC:

$$\begin{aligned} \text{AUPRC} &= \frac{1 - \text{Recall} + \text{Precision}}{2} \\ &= \frac{1 - \text{Sensitivity} + \text{Precision}}{2}. \end{aligned} \tag{3.27}$$

On the other hand, notice that Matthew's correlation takes into account both recall and precision while correcting for bias and prevalence at chance level unlike AUC or AUPRC.

4.0 THE CAUSAL FRAMEWORK

The aforementioned Bayesian networks, dynamic Bayesian networks and SEM-IEs describe three causal frameworks. I believe these frameworks are excellent for studying causality, but I also believe that they have several weaknesses when describing real causal processes as argued in Section 4.1. As a result, I describe a new causal framework in Sections 4.2 and 4.3 as well as its associated sampling process in Section 4.4, which I believe are more realistic. I finally characterize the conditional independence properties present in the induced distribution in Sections 4.5 and 4.6.

4.1 WEAKNESSES OF PREVIOUS CAUSAL FRAMEWORKS

Bayesian networks, dynamic Bayesian networks and SEM-IEs describe three frameworks for representing causal processes. While each framework carries its strengths, each framework also carries weaknesses which raise questions about its applicability to real situations. I list some of the weaknesses below:

1. Many causal processes in nature appear to contain feedback loops, which we cannot model with Bayesian networks.
2. Time may be continuous rather than discrete, so dynamic Bayesian networks may only provide an approximation of a continuous time causal process. Ideally, we would like to model a continuous time causal process with continuous time rather than discrete time.
3. Obtaining measurements from time series variables appears unrealistic in many contexts, since scientists often cannot measure random variables at the exact same time point in

the causal process for each sample. For example, biologists may measure the protein levels in liver cells, but each cell may lie at a different time point in the causal process.

4. Even if one could measure random variables at the same time points in special cases, multiple measurements of the random variables at different time points may not be possible due to technical difficulties, ethical issues, or monetary constraints. We also need to acknowledge that we have more data with single measurements than data with multiple measurements.
5. We often cannot obtain sample trajectories as required for learning CTBNs [Nodelman et al., 2003, 2005]. Moreover, we often must deal with non-parametric continuous random variables which CTBNs currently cannot handle.
6. I do not believe that nature executes the fixed point method when sampling from SEM-IEs containing cycles or chain graphs because I am hesitant to assume that successively applied functional transformations in nature do not contain any noise.
7. Even if we can represent a real causal process using an SEM-IE and apply the fixed point method, we may have non-linear structural equations, so the global directed Markov property may not hold when cycles exist.

In this chapter, I seek to alleviate the aforementioned difficulties by developing a new, modified causal framework. This framework in turn will help in the rigorous development of an algorithm for causal discovery under more realistic conditions.

4.2 CONTINUOUS TIME STOCHASTIC PROCESSES WITH JUMP POINTS

I now proceed to define a Continuous time stochastic process with Jump points (CJ). I first take time as naturally continuous and therefore consider a continuous time stochastic process by setting $\Theta = [0, \infty)$ in Definition 3; note that time point zero denotes the beginning of a well-defined stochastic process, but not necessarily the "beginning of time." Now let $\widetilde{\boldsymbol{X}} = \Big\{\{X_1^t | t \in \Theta\}, \{X_2^t | t \in \Theta\}, \dots\Big\}$ and $X_i^\Theta = \{X_i^t | t \in \Theta\} \in \widetilde{\boldsymbol{X}}$. Then, for each $X_i^\Theta \in \widetilde{\boldsymbol{X}}$, I introduce a set of fixed *jump points* $J_i = \{J_{i,1} = 0, J_{i,2}, J_{i,3}, \dots\}$ such that $0 < J_{i,2} < J_{i,3} <$

... and $J_{i,k} \in \Theta, \forall k$. I require that:

$$X_i^{\tau+J_{i,k}} = X_i^{J_{i,k+1}}, \tag{4.1}$$

where $\tau \in [0, J_{i,k+1} - J_{i,k})$, and I refer to the interval $[0, J_{i,k+1} - J_{i,k}) = I_i^{J_{i,k}}$ as the *holding interval* of X_i at jump point $J_{i,k}$. In other words, the value of X_i does not change within the holding interval between the jump points. Due to the equivalence in (4.1), I will often refer to $X_i^{J_{i,k}}$ as simply X_i^t at some $t \in [J_{i,k}, J_{i,k+1})$.

A simple way of illustrating a CJ involves first creating an axis representing time and then placing a vertex for each random variable at each of its jump points. I provide an example of an illustration of a CJ with two variables X_1^Θ and X_2^Θ in Figure 4.1a. Notice that the CJ has well-defined random variables at any time point starting from time point zero, since the set of jump points for each variable must include a jump point at time zero. Moreover, the values of each variable remain the same within the holding intervals of each variable, as shown in the middle portion of Figure 4.1b. Jump points therefore ensure that scientists have some finite amount of time to measure each X_i^t, as measuring variables naturally takes time in biomedicine (and most other fields in science as far as I am aware). For example, subjects need time to complete questionnaires and antibodies need time to bind.

4.3 ADDING THE MARKOV ASSUMPTION

I now proceed to make the CJ Markov, and I choose to represent the CJ's Markovian nature using a DAG over $\boldsymbol{X}^\Theta$. I create the DAG by drawing directed edges between the vertices in the CJ. I assume that all causal relationships are either instantaneous (i.e., take zero time to complete) or non-instantaneous (i.e., take some positive finite amount of time to complete).[1] Thus, for each variable X_i at jump point $J_{i,k}$, I consider its parent set $\boldsymbol{Pa}(X_i^{J_{i,k}})$, where $X_a^u \in \boldsymbol{Pa}(X_i^{J_{i,k}})$ if and only if $u \leq J_{i,k}$, $u \in J_a$, and X_a^u has a directed edge towards $X_i^{J_{i,k}}$ in the DAG. I will call this DAG $\mathbb{G}$ the *CMJ-DAG* from here on. I have provided an example

[1] As commonly assumed in the existing literature, I do not allow causal relations to take some negative finite amount of time to complete and thus be directed backwards in time.

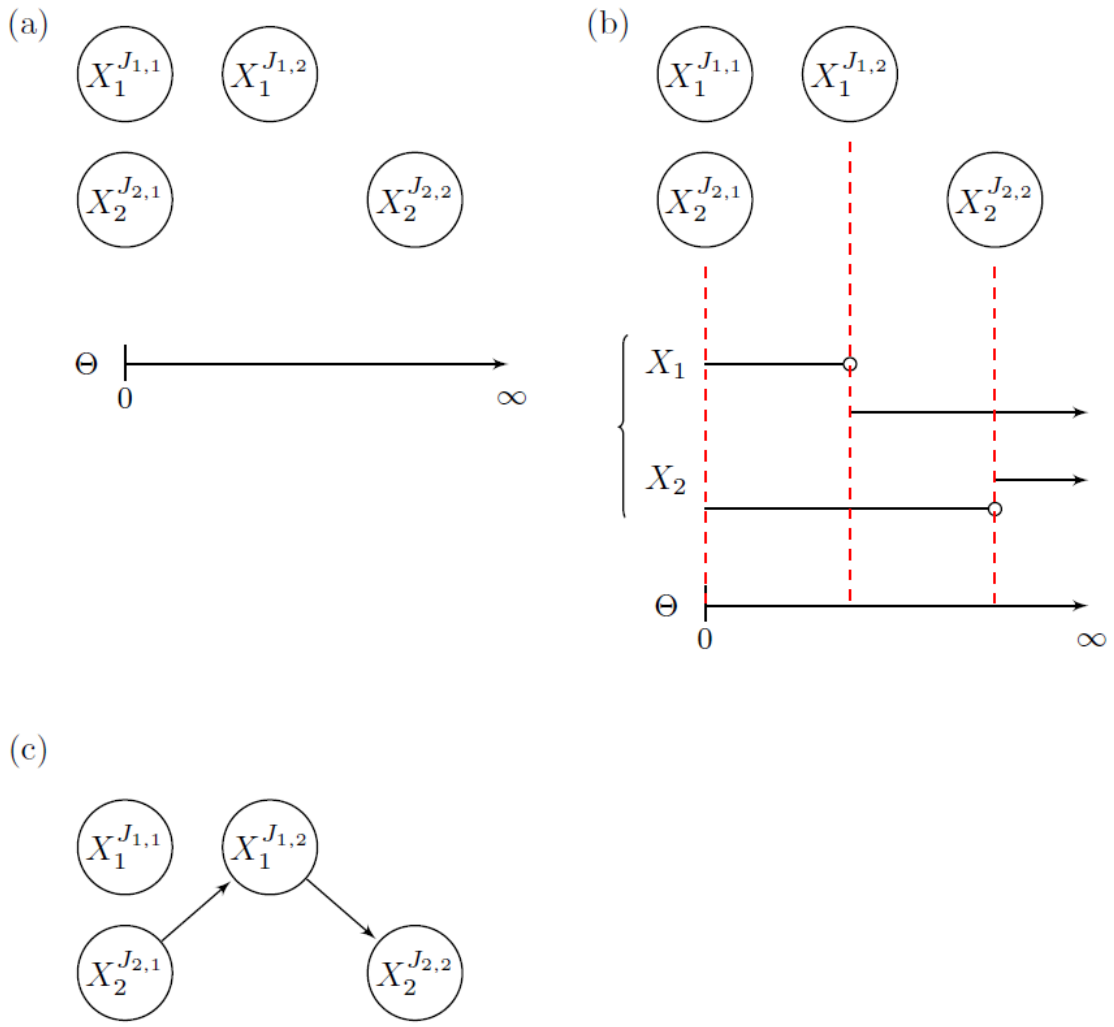


Figure 4.1: (a) An illustration of a CJ with two variables X_1^{Θ} and X_2^{Θ}. The bottom arrow labeled with Θ denotes continuous time. The top portion of the figure shows the two variables at two of their jump points – X_1^{Θ} at its first and second jump points denoted as $X_1^{J_{1,1}}$ and $X_1^{J_{1,2}}$ respectively, and X_2^{Θ} at its first and second jump points denoted as $X_2^{J_{2,1}}$ and $X_2^{J_{2,2}}$ respectively. (b) The CJ in (a) but with an additional middle portion contained within a bracket showing the evolution of the values of the variables across continuous time. The red dashed lines track each of the jump points across the entire figure. (c) An example of a CMJ-DAG using the CJ in (a).

in Figure 4.1c; notice that every directed edge originates from and connects to a random variable at its jump point in the CJ.

Take note that I do allow instantaneous causal effects from parents to their children due to the non-strict inequality $u \leq J_{i,k}$. I however only consider CJs with DAGs and therefore will assume *acyclic* contemporaneous sub-graphs throughout this thesis; in fact, I have not discovered any way to realistically justify instantaneous cyclic relations at the time of writing this thesis. We can of course consider the fixed point in the fixed point method as a type of instantaneous cyclic causal relation, but this method appears unrealistic to me because I am hesitant to assume that successively applied functional transformations in nature do not contain any noise. I however can think of a realistic situation where instantaneous acyclic causal relations exist. For example, scientists can introduce instantaneous causal effects in medicine by introducing summary variables into a dataset, since a summary variable exists instantaneously once all of its component variables exist. As a specific clinical example, physicians may include the composite Mini-Mental Status Exam score as well as some of the test's memory component scores in an Alzheimer's disease dataset.

Now I say that a finite dimensional distribution $\mathbb{P}_{\boldsymbol{X}^{\Theta^T}}$ over $\boldsymbol{X}^{\Theta^T} = \{\boldsymbol{X}^t | t \in \Theta^T\}$ where $\Theta^T \subseteq [0, r_2], r_2 \in \mathbb{R}_{\geq 0}$ with density $f(\boldsymbol{X}^{\Theta^T})$ satisfies the *Markov property* with respect to a CMJ-DAG associated with a CJ if and only if:

$$f(\boldsymbol{X}^{\Theta^T}) = \prod_{i=1}^{p} \prod_{k=1}^{|J_i|} f(X_i^{J_{i,k}} | \boldsymbol{Pa}(X_i^{J_{i,k}})), \tag{4.2}$$

where $J_{i,k} \in \Theta^T, \forall i, k$.

I now define a continuous time Markov process with jump points (CMJ) over Θ^T as follows:

Definition 4. *A CMJ over Θ^T is a CJ with a finite dimensional distribution over $\boldsymbol{X}^{\Theta^T}$ that satisfies the global directed Markov property with respect to the CMJ-DAG $\mathbb{G}$.*

More compactly, we can consider a CMJ as the double $(\mathbb{G}, \mathbb{P}_{\boldsymbol{X}^{\Theta^T}})$, where $\mathbb{G}$ denotes a CMJ-DAG and $\mathbb{P}_{\boldsymbol{X}^{\Theta^T}}$ the finite dimensional distribution over $\boldsymbol{X}^{\Theta^T}$ that satisfies the global directed Markov property with respect to $\mathbb{G}$.

Notice that the CMJ is a generalization of the BN, since we can interpret a BN as a CMJ where $\boldsymbol{X}^{\Theta^T}$ has one jump point at time point zero. The CMJ gives the additional flexibility of allowing the distribution over $\boldsymbol{X}^{\Theta^T}$ to change over continuous time. I say that there exists a *feedback loop* involving $X_i^{\Theta^T}$ in a CMJ-DAG if and only if there exists a directed path from $X_i^{J_{i,a}}$ to $X_i^{J_{i,b}}$ in the CMJ-DAG such that $J_{i,a}, J_{i,b} \in \Theta^T$. The CMJ thus also gives us the ability to represent a "non-stationary cyclic causal process" in a well-defined form. In this sense, the CMJ is a natural generalization of the BN to continuous time.

4.3.1 Criticisms of the CMJ

Whenever we propose a new framework for causality, we should analyze both its strengths and weaknesses. I have thus far only discussed the strengths of the CMJ model. In this section, I discuss some concerning properties about the CMJ.

Some of my colleagues have suggested incorporating *stochastic* rather than fixed jump points into the CMJ. Indeed, many authors publishing in the stochastic process literature have described processes involving stochastic jump points (e.g., Poisson processes [Stoyan et al., 1987]). Note that nature may determine the jump points of the CMJ by a stochastic mechanism; I place no restriction on how nature creates a CMJ. However, notice that the jump points must ultimately be fixed once nature determines their times. Recall also that data used to infer causation always contains measurements of random variables realized in the past whose jump points therefore have also already been realized. Thus, no difference exists between a CMJ with fixed jump points compared to a CMJ with stochastic jump points once nature has instantiated the jump points. In other words, we can consider a CMJ with fixed jump points and a CMJ with stochastic jump points as equivalent in the context of causal discovery, so long as we only consider independent samples from one underlying CMJ model.

Naturally then one may wonder whether it is appropriate to have one underlying CMJ model. Indeed, some of my colleagues have suggested that, while one underlying causal process may exist, the CMJ may actually unfold at different speeds for each sample. For example, time $t = 10$ in the CMJ of the first sample may correspond to time $t = 15$ (instead

of $t = 10$) in the CMJ of the second sample, because the second sample's CMJ proceeds at a faster pace. This interpretation of varying speeds is still congruent to the proposed idea of a single underlying CMJ model, so long as we adjust our interpretation of T. Instead of interpreting a time point t as the amount of time passed from time point zero in any sample, we interpret t as the amount of time passed from time point zero *for each sample*. The proposed CMJ model therefore easily incorporates the varying speed suggestion as well.

The CMJ model however does not incorporate all suggestions. The aforementioned concern considers CMJs with different speeds per sample, but we can go even further and consider an arbitrarily different CMJ per sample. We may for example associate each CMJ with a different CMJ-DAG or a different ordering of jump points. In cases such as these, we cannot discover causal structure across the possibly infinite set of CMJs without imposing regularity conditions. For example, we may impose the condition that there only exists a finite number of possible CMJs. We may also require that every CMJ-DAG must share certain d-connection relations.

In this thesis, I will assume the existence of one underlying CMJ model just like how we assume the existence of one underlying BN model when performing acyclic causal discovery. I will leave extensions to sets of CMJ models open to future work.

4.4 MIXING THE DISTRIBUTIONS OF A CMJ

Recall that we have defined the finite dimensional joint distribution of a CMJ over $\boldsymbol{X}^{\Theta^T}$. We can of course consider inferring a CMJ, if we can collect samples from the joint distribution of the CMJ by sampling at many known time points with small enough time intervals. Scientists typically collect such data in medicine when they can kick-start a causal process at a pre-specified time and therefore specify time point zero in the causal process. For example, neuroscientists may sample the blood-oxygen level dependent (BOLD) signal from the brain after inducing a causal process with a specific task in an fMRI machine [Ramsey et al., 2010, Smith et al., 2011, Iyer et al., 2013]. Obtaining time series data therefore often requires setting up controlled environments just like with experimental data and necessitates

active participation during the data collection process.[2]

In this thesis, I will focus instead on discovering causal structure by *passively* observing the variables in their natural environment, where we usually cannot align the samples according to time. For instance, a medical investigator may recruit individuals with a specific disease and measure their physical characteristics every week; observe that the individuals have the same disease, but some of them may be late in the disease process while others may be early. Indeed, data generated from a CMJ should naturally come from different time points, since scientists usually have no way of identifying the time point for each sample. Even if one attempts to align the samples in time using proxy variables such as age or first physician's visit, the proxy variables may only allow an approximate alignment.

I therefore find it reasonable to treat time T as a random variable mapping onto Θ^T with Borel sigma algebra $\mathcal{B}_{\Theta^T}$ and distribution $\mathbb{P}_T$. A typical sampling process involves the following procedure. First, we draw the value of T according to $\mathbb{P}_T$; denote t as the value drawn. Next, we draw a sample of $\boldsymbol{X}$ according to $\mathbb{P}_{\boldsymbol{X}^t}$, the finite dimensional distribution over $\boldsymbol{X}$ at time point t. I can therefore represent the resulting density over $\boldsymbol{X}$ as a mixture of densities over $\boldsymbol{X}$ at different time points:

$$f_m(\boldsymbol{X}) \stackrel{\text{def}}{=} \int_{\Theta^T} f_t(\boldsymbol{X}) \, d\mathbb{P}_T, \tag{4.3}$$

where $f_t(\boldsymbol{X})$ refers to the density of the probability distribution $\mathbb{P}_{\boldsymbol{X}^t}$; I will sometimes equivalently write $f_t(\boldsymbol{X})$ as $f(\boldsymbol{X}^t)$, when I want to place emphasis on the variables at time point t rather than on the density. Let $\mathbb{P}_{\boldsymbol{X}^m}$ similarly refer to the probability distribution of the *mixture density* $f_m(\boldsymbol{X})$.

Consider the dataset in Table 4.1 as an example of data generated according to $f_m(\boldsymbol{X})$. This dataset has at least three samples, which I created by sampling the CMJ in Figure 4.2 uniformly between $[0.05, 0.35] \cup [0.55, 0.85]$. Note that each sample in Table 4.1 contains time point information but, in practice, an investigator usually also cannot observe the time points of each sample. As a result, scientists often cannot (1) convert this type of data into time series data by aligning the samples according to time, or (2) use time as an additional

[2]Scientists may also get lucky with a "natural change" in the environment, such as a new government policy, which initiates a causal process.

surrogate variable [Spirtes, 1994, Zhang et al., 2015]. From here on, I will call a dataset created from sampling a CMJ at random time points as *mixture data*. In my opinion, nearly all real datasets collected by passive observation are in fact mixture datasets.

The random samplings in time unfortunately introduce several theoretical concerns when we try to use mixture data to reconstruct a CMJ-DAG. Intuitively, samples drawn according to $f_m(\boldsymbol{X})$ may contain less information about the causal structure at time point t than samples drawn according to $f_t(\boldsymbol{X})$. Moreover, the mixing process may introduce some unwanted dependencies which may complicate discovery of the CMJ-DAG. In the next section, I aim to better understand these concerns by characterizing the properties of the joint mixture density in two types of CMJs.

4.5 STATIONARY CMJS

I will consider two types of CMJs in this thesis: stationary and non-stationary. Let $S(f(T))$ denote the support of $f(T)$, the density of $\mathbb{P}_T$. I say that a CMJ is *stationary* over $S(f(T))$ if and only if $\forall t, t' \in S(f(T)), \mathbb{P}_{\boldsymbol{X}^t} = \mathbb{P}_{\boldsymbol{X}^{t'}}$. Equivalently, a CMJ is *non-stationary* over $S(f(T))$ if and only if $\exists t, t' \in S(f(T))$ such that $\mathbb{P}_{\boldsymbol{X}^t} \neq \mathbb{P}_{\boldsymbol{X}^{t'}}$.

The stationary scenario is actually trivial, because we have equality in distribution over time. Recall that the CMJ-DAG is acyclic over $\boldsymbol{X}^\Theta$ and therefore also acyclic over $\boldsymbol{X}^{[0,t_1]}$, where t_1 denotes the earliest time point in $S(f(T))$. The causal discovery task in the stationary situation thus equates to the usual acyclic causal discovery task using the distribution $\mathbb{P}_{\boldsymbol{X}^{t_1}}$. Under d-separation faithfulness, an existing algorithm such as FCI applies here even when feedback loops exist, so long as we understand that the algorithm only has access to samples drawn according to $\mathbb{P}_{\boldsymbol{X}^{t_1}}$ and therefore only attempts to recover the CMJ-DAG at time t_1.

I provide two examples of stationary CMJs in Figure 4.3a. The first example depicts a CMJ where we have equality in random variables over $S(f(T))$. On the other hand, the CMJ in Figure 4.3b allows the values of X_2 to change over time. In either case, sampling from some mixture of $\{\mathbb{P}_{\boldsymbol{X}^t} : t \in S(f(T))\}$ is equivalent to sampling from $\mathbb{P}_{\boldsymbol{X}^{t_1}}$.

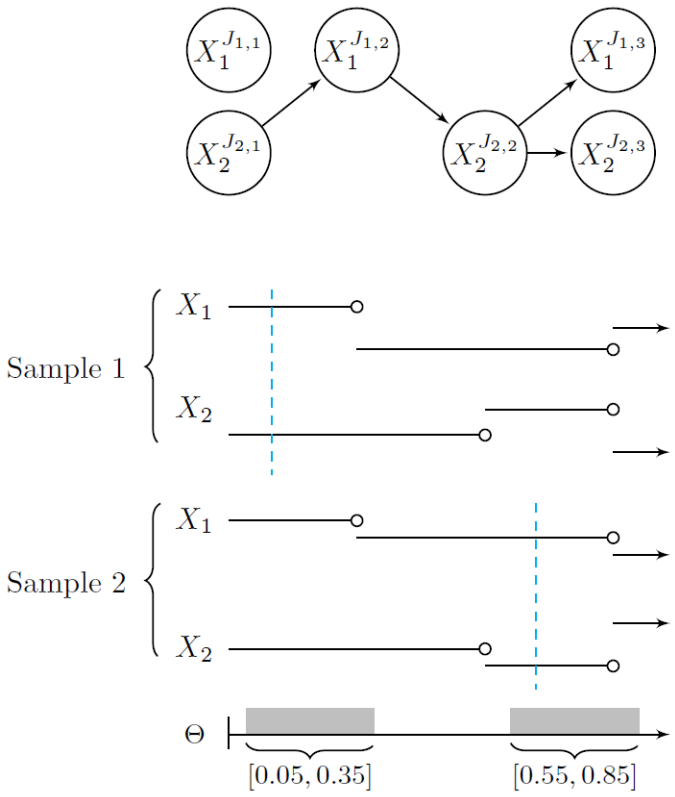


Figure 4.2: A sampling process from a CMJ. Here, we draw time uniformly from $[0.05, 0.35] \cup [0.55, 0.85]$. The blue lines denote the time points of the first and second samples. The sampling depicted in this figure thus creates two of the samples in the dataset shown in Table 4.1.

Sample	X_1	X_2
1	$0.53^{0.16}$	$-1.40^{0.16}$
2	$0.36^{0.61}$	$-1.09^{0.61}$
$\vdots$	$\vdots$	

Table 4.1: An example of mixture data generated by the sampling process in Figure 4.2. Superscripts denote the time points of X_1 and X_2 which are typically hidden.

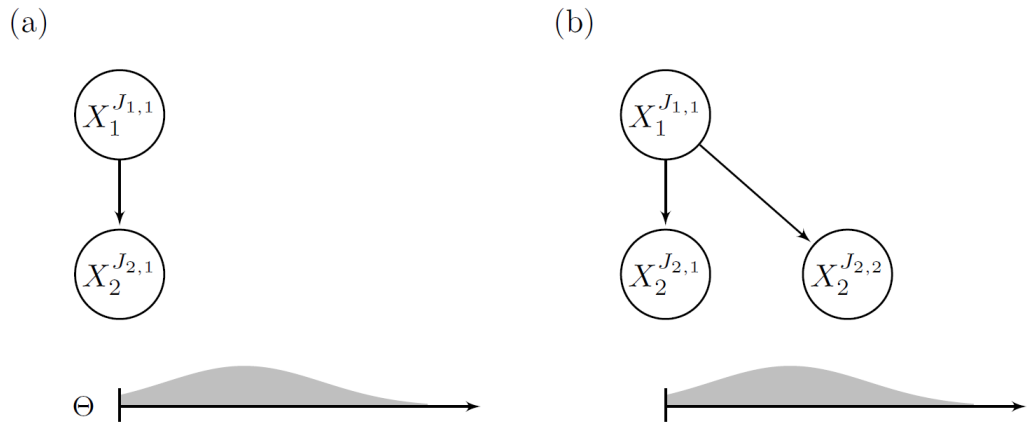


Figure 4.3: Two examples of stationary CMJs. (a) depicts a CMJ where the random variables do not change their values over time, while (b) depicts a CMJ where X_2 can change its value.

Note that the definition of stationarity used here differs from the definition of *equilibrium* used in SEM-IEs. An SEM-IE achieves equilibrium if and only if all of its structural equations are simultaneously satisfied. In the cyclic case, this means that an SEM-IE can admit instantaneous cyclic causal relations in the sense that if X_i is a cause of X_j and vice versa, then X_i causes X_j at the same time that X_j causes X_i. Recall that we do not allow instantaneous cycles in the CMJ-DAG. Thus, stationarity as used in this thesis simply means that the distribution $\mathbb{P}_{\boldsymbol{X}^{t_1}}$ does not change over $S(f(T))$. We therefore do not run into the causal interpretability problems of equilibriated or stationary SEM-IEs (i.e., on the commutability of the Do and Equilibration operators; see [Dash, 2005] for details).

4.6 NON-STATIONARY CMJS

I now analyze the properties of non-stationary CMJs in detail. I have divided the analysis into four separate subsections, since this scenario involves significantly more complex arguments.

4.6.1 Conditional Independence Properties

Let us first consider conditional independence properties according to $f_m(\boldsymbol{X})$. Let $\boldsymbol{X} = \{\boldsymbol{O} \cup \boldsymbol{L} \cup \boldsymbol{S}\}$ and define:

$$f_m(\boldsymbol{U}|\boldsymbol{W} = \boldsymbol{w}, \boldsymbol{S} = \boldsymbol{s}) \stackrel{\text{def}}{=} \int_{\Theta^T} f(\boldsymbol{U}^t|\boldsymbol{W}^t = \boldsymbol{w}, \boldsymbol{S}^t = \boldsymbol{s})\ d\mathbb{P}_T(t), \tag{4.4}$$

where $\boldsymbol{U}, \boldsymbol{W} \subseteq \boldsymbol{O}$.

I say $f(\boldsymbol{U}^t|\boldsymbol{W}^t, \boldsymbol{S}^t) = f(\boldsymbol{U}^{t'}|\boldsymbol{W}^{t'}, \boldsymbol{S}^{t'})$ if and only if $f(\boldsymbol{U}^t|\boldsymbol{W}^t = \boldsymbol{w}, \boldsymbol{S}^t = \boldsymbol{s}) = f(\boldsymbol{U}^{t'}|\boldsymbol{W}^{t'} = \boldsymbol{w}, \boldsymbol{S}^{t'} = \boldsymbol{s})$ for any $\boldsymbol{w}, \boldsymbol{s}$ such that $f(\boldsymbol{W}^t = \boldsymbol{w}, \boldsymbol{S}^t = \boldsymbol{s}) > 0$ and $f(\boldsymbol{W}^{t'} = \boldsymbol{w}, \boldsymbol{S}^{t'} = \boldsymbol{s}) > 0$. I will use the phrase "*indexing variable* of $f(\boldsymbol{U}|\boldsymbol{W}, \boldsymbol{S})$" to refer to a variable denoted as $\mathcal{I}_{\boldsymbol{U}|\boldsymbol{W},\boldsymbol{S}}$ which indexes every unique conditional density in the set $\{f(\boldsymbol{U}^t|\boldsymbol{W}^t, \boldsymbol{S}^t) : t \in S(f(T))\}$ using each unique density's earliest time point on $S(f(T))$. Let $f(\boldsymbol{U}|\boldsymbol{W}, \boldsymbol{S}, T = t) \stackrel{\text{def}}{=} f(\boldsymbol{U}^t|\boldsymbol{W}^t, \boldsymbol{S}^t)$; thus, $f(\boldsymbol{U}|\boldsymbol{W}, \boldsymbol{S}, \mathcal{I}_{\boldsymbol{U}|\boldsymbol{W},\boldsymbol{S}} = a) \neq f(\boldsymbol{U}|\boldsymbol{W}, \boldsymbol{S}, \mathcal{I}_{\boldsymbol{U}|\boldsymbol{W},\boldsymbol{S}} = b), \forall a \neq b$ which are the earliest time points of the unique densities on $S(f(T))$. For shorthand, I will sometimes write $f(\boldsymbol{U}|\boldsymbol{W}, \boldsymbol{S}, \mathcal{I}_{\boldsymbol{U}|\boldsymbol{W},\boldsymbol{S}} = a)$ as $f(\boldsymbol{U}^a|\boldsymbol{W}^a, \boldsymbol{S}^a)$, since a is a time point.

We can more specifically view the indexing variable as a random variable with codomain $\Theta_{\mathcal{I}_{\boldsymbol{U}|\boldsymbol{W},\boldsymbol{S}}}$ and probability distribution:

$$\mathbb{P}_{\mathcal{I}_{\boldsymbol{U}|\boldsymbol{W},\boldsymbol{S}}}(a) = \int_{\Delta^a} d\mathbb{P}_T, \tag{4.5}$$

where $\Delta^a \in \mathcal{B}_{S(f(T))}$ denotes the largest Borel set of time points on $S(f(T))$ such that $f(\boldsymbol{U}^a|\boldsymbol{W}^a, \boldsymbol{S}^a)$ remains unchanged; that is, $f(\boldsymbol{U}^t|\boldsymbol{W}^t, \boldsymbol{S}^t) = f(\boldsymbol{U}^a|\boldsymbol{W}^a, \boldsymbol{S}^a)$ for any $t, a \in \Delta^a, a \leq t$. Note that $|\Theta_{\mathcal{I}_{\boldsymbol{U}|\boldsymbol{W},\boldsymbol{S}}}| \geq 1$. I say that the set $\{f(\boldsymbol{U}^t|\boldsymbol{W}^t, \boldsymbol{S}^t) : t \in S(f(T))\}$ is non-stationary over $S(f(T))$ if and only if $|\Theta_{\mathcal{I}_{\boldsymbol{U}|\boldsymbol{W},\boldsymbol{S}}}| > 1$. For shorthand, I will say that $f(\boldsymbol{U}|\boldsymbol{W}, \boldsymbol{S})$ is stationary (or non-stationary) when in fact $\{f(\boldsymbol{U}^t|\boldsymbol{W}^t, \boldsymbol{S}^t) : t \in S(f(T))\}$ is stationary (or non-stationary).[3]

Notice that we can more specifically view $\mathcal{I}_{\boldsymbol{U}|\boldsymbol{W},\boldsymbol{S}}$ as an *index of density parameters* for the conditional density $f(\boldsymbol{U}|\boldsymbol{W}, \boldsymbol{S})$ rather than an index of time. That is, we consider a

[3]Note that we can state the definition of stationarity of a conditional density in a fashion similar to the definition of stationarity of a CMJ over $S(f(T))$; that is, $f(\boldsymbol{U}|\boldsymbol{W}, \boldsymbol{S})$ is stationary over $S(f(T))$ if and only if $f(\boldsymbol{U}^t|\boldsymbol{W}^t, \boldsymbol{S}^t) = f(\boldsymbol{U}^{t'}|\boldsymbol{W}^{t'}, \boldsymbol{S}^{t'}), \forall t, t' \in S(f(T))$. This definition however is more cumbersome for the purposes of my argument, so I will use the previous definition when referring to particular conditional densities rather than the entire CMJ.

family of densities $\{f(\boldsymbol{U}|\boldsymbol{W},\boldsymbol{S},\boldsymbol{\lambda}) : \boldsymbol{\lambda} \in \Lambda\}$, where $\Theta_{\mathcal{I}_{\boldsymbol{U}|\boldsymbol{W},\boldsymbol{S}}}$ indexes the set Λ. We can thus write $f(\boldsymbol{U}|\boldsymbol{W},\boldsymbol{S},\mathcal{I}_{\boldsymbol{U}|\boldsymbol{W},\boldsymbol{S}} = a) = f(\boldsymbol{U}|\boldsymbol{W},\boldsymbol{S},\boldsymbol{\lambda}_a), \forall a \in \Theta_{\mathcal{I}_{\boldsymbol{U}|\boldsymbol{W},\boldsymbol{S}}}$.

I write $O_i \perp\!\!\!\perp O_j|(\boldsymbol{W},\boldsymbol{S})$ whenever conditional independence holds in the mixture density; in other words, $O_i \perp\!\!\!\perp O_j|(\boldsymbol{W},\boldsymbol{S})$ if and only if $f_m(O_i, O_j|\boldsymbol{W},\boldsymbol{S}) = f_m(O_i|\boldsymbol{W},\boldsymbol{S})f_m(O_j|\boldsymbol{W},\boldsymbol{S})$. We now have the following result:

Proposition 3. *For any O_i, O_j and $\boldsymbol{W} \subseteq \boldsymbol{O} \setminus \{O_i, O_j\}$, if $O_i \perp\!\!\!\perp O_j|(\boldsymbol{W},\boldsymbol{S},\boldsymbol{\lambda}_{\mathcal{I}_{O_iO_j|\boldsymbol{W},\boldsymbol{S}}})$ and $\boldsymbol{\lambda}_{\mathcal{I}_{O_i|\boldsymbol{W},\boldsymbol{S}}} \perp\!\!\!\perp \boldsymbol{\lambda}_{\mathcal{I}_{O_j|\boldsymbol{W},\boldsymbol{S}}}$, then $O_i \perp\!\!\!\perp O_j|(\boldsymbol{W},\boldsymbol{S})$.*

Proof. We may write the following equalities for the forward direction:

$$
\begin{aligned}
& f_m(O_i, O_j|\boldsymbol{W},\boldsymbol{S}) \\
&= \int f(O_i, O_j|\boldsymbol{W},\boldsymbol{S},\boldsymbol{\lambda}_{\mathcal{I}_{O_iO_j|\boldsymbol{W},\boldsymbol{S}}})\, d\mathbb{P}_{\boldsymbol{\lambda}_{\mathcal{I}_{O_iO_j|\boldsymbol{W},\boldsymbol{S}}}} \\
&= \int f(O_i|\boldsymbol{W},\boldsymbol{S},\boldsymbol{\lambda}_{\mathcal{I}_{O_iO_j|\boldsymbol{W},\boldsymbol{S}}}) f(O_j|\boldsymbol{W},\boldsymbol{S},\boldsymbol{\lambda}_{\mathcal{I}_{O_iO_j|\boldsymbol{W},\boldsymbol{S}}})\, d\mathbb{P}_{\boldsymbol{\lambda}_{\mathcal{I}_{O_iO_j|\boldsymbol{W},\boldsymbol{S}}}} \\
&= \int f(O_i|\boldsymbol{W},\boldsymbol{S},\boldsymbol{\lambda}_{\mathcal{I}_{O_i|\boldsymbol{W},\boldsymbol{S}}}) f(O_j|\boldsymbol{W},\boldsymbol{S},\boldsymbol{\lambda}_{\mathcal{I}_{O_j|\boldsymbol{W},\boldsymbol{S}}})\, d\mathbb{P}_{\boldsymbol{\lambda}_{\mathcal{I}_{O_i|\boldsymbol{W},\boldsymbol{S}}}\boldsymbol{\lambda}_{\mathcal{I}_{O_j|\boldsymbol{W},\boldsymbol{S}}}} \\
&= \int f(O_i|\boldsymbol{W},\boldsymbol{S},\boldsymbol{\lambda}_{\mathcal{I}_{O_i|\boldsymbol{W},\boldsymbol{S}}})\, d\mathbb{P}_{\boldsymbol{\lambda}_{\mathcal{I}_{O_i|\boldsymbol{W},\boldsymbol{S}}}} \int f(O_j|\boldsymbol{W},\boldsymbol{S},\boldsymbol{\lambda}_{\mathcal{I}_{O_j|\boldsymbol{W},\boldsymbol{S}}})\, d\mathbb{P}_{\boldsymbol{\lambda}_{\mathcal{I}_{O_j|\boldsymbol{W},\boldsymbol{S}}}} \\
&= f_m(O_i|\boldsymbol{W},\boldsymbol{S}) f_m(O_j|\boldsymbol{W},\boldsymbol{S}).
\end{aligned}
\tag{4.6}
$$

The second equality follows by the first sufficient condition and the fourth equality by the second sufficient condition.

□

4.6.2 Conditional Dependence Across Time

The contrapositive of Proposition 3 allows us to make important conclusions about conditional dependence as a function of time when conditional dependence holds in the mixture density. We may re-write the contrapositive directly in terms of time as follows:

Lemma 1. *If $O_i \not\perp\!\!\!\perp O_j|(\boldsymbol{W},\boldsymbol{S})$, then $O_i^t \not\perp\!\!\!\perp O_j^t|(\boldsymbol{W}^t,\boldsymbol{S}^t)$ at some $t \in S(f(T))$ or we have parameter dependence: $\boldsymbol{\lambda}_{\mathcal{I}_{O_i|\boldsymbol{W},\boldsymbol{S}}} \not\perp\!\!\!\perp \boldsymbol{\lambda}_{\mathcal{I}_{O_j|\boldsymbol{W},\boldsymbol{S}}}$.*

Here, I write $O_i^t \perp\!\!\!\perp O_j^t | (\boldsymbol{W}^t, \boldsymbol{S}^t)$ if and only if $f(O_i^t, O_j^t | \boldsymbol{W}^t, \boldsymbol{S}^t) = f(O_i^t | \boldsymbol{W}^t, \boldsymbol{S}^t)\, f(O_j^t | \boldsymbol{W}^t, \boldsymbol{S}^t)$. Therefore, equivalently, $O_i^t \not\perp\!\!\!\perp O_j^t | (\boldsymbol{W}^t, \boldsymbol{S}^t)$ if and only if $f(O_i^t, O_j^t | \boldsymbol{W}^t, \boldsymbol{S}^t) \neq f(O_i^t | \boldsymbol{W}^t, \boldsymbol{S}^t)\, f(O_j^t | \boldsymbol{W}^t, \boldsymbol{S}^t)$.

We would like to claim that if $O_i \not\perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S})$, then $O_i^t \not\perp\!\!\!\perp O_j^t | (\boldsymbol{W}^t, \boldsymbol{S}^t)$ at some $t \in S(f(T))$. However, the parameter dependence $\boldsymbol{\lambda}_{\mathcal{I}_{O_i|\boldsymbol{W},\boldsymbol{S}}} \not\perp\!\!\!\perp \boldsymbol{\lambda}_{\mathcal{I}_{O_j|\boldsymbol{W},\boldsymbol{S}}}$ introduces a problem; even if we have $O_i \not\perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S})$, we might still have $O_i^t \perp\!\!\!\perp O_j^t | (\boldsymbol{W}^t, \boldsymbol{S}^t)$ at all $t \in S(f(T))$ with $\boldsymbol{\lambda}_{\mathcal{I}_{O_i|\boldsymbol{W},\boldsymbol{S}}} \not\perp\!\!\!\perp \boldsymbol{\lambda}_{\mathcal{I}_{O_j|\boldsymbol{W},\boldsymbol{S}}}$.

We can fortunately ensure that we have $O_i^t \not\perp\!\!\!\perp O_j^t | (\boldsymbol{W}^t, \boldsymbol{S}^t)$ at some $t \in S(f(T))$ by checking additional conditions. The following two lemmas serve as useful tools for dissecting out conditional dependencies by utilizing conditional mixture modeling:

Lemma 2. *If (1) $O_i \not\perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S})$ and (2) $f(O_i, O_j | \boldsymbol{W}, \boldsymbol{S})$ is stationary, then $O_i^t \not\perp\!\!\!\perp O_j^t | (\boldsymbol{W}^t, \boldsymbol{S}^t)$ at all $t \in S(f(T))$.*

Proof. If $O_i \not\perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S})$, then (1) $O_i \not\perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S}, \boldsymbol{\lambda}_{\mathcal{I}_{O_i O_j|\boldsymbol{W},\boldsymbol{S}}})$ or (2) $\boldsymbol{\lambda}_{\mathcal{I}_{O_i|\boldsymbol{W},\boldsymbol{S}}} \not\perp\!\!\!\perp \boldsymbol{\lambda}_{\mathcal{I}_{O_j|\boldsymbol{W},\boldsymbol{S}}}$ (or both) holds by Proposition 3. We also know that $f(O_i, O_j | \boldsymbol{W}, \boldsymbol{S})$ is stationary, so we must more specifically have $\boldsymbol{\lambda}_{\mathcal{I}_{O_i|\boldsymbol{W},\boldsymbol{S}}} \perp\!\!\!\perp \boldsymbol{\lambda}_{\mathcal{I}_{O_j|\boldsymbol{W},\boldsymbol{S}}}$ because $\boldsymbol{\lambda}_{\mathcal{I}_{O_i O_j|\boldsymbol{W},\boldsymbol{S}}}$ is constant (call it c). We conclude that $O_i \not\perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S}, \boldsymbol{\lambda}_{\mathcal{I}_{O_i O_j|\boldsymbol{W},\boldsymbol{S}}} = c)$, so $O_i^t \not\perp\!\!\!\perp O_j^t | (\boldsymbol{W}^t, \boldsymbol{S}^t)$ at all $t \in S(f(T))$. □

We also have the following:

Lemma 3. *If (1) $O_i \not\perp\!\!\!\perp O_j | \boldsymbol{W}, \boldsymbol{S}$ and (2) at least one member of $\{f(O_i | \boldsymbol{W}, \boldsymbol{S}), f(O_j | \boldsymbol{W}, \boldsymbol{S})\}$ is stationary, then $O_i^t \not\perp\!\!\!\perp O_j^t | (\boldsymbol{W}^t, \boldsymbol{S}^t)$ at some $t \in S(f(T))$.*

Proof. If $O_i \not\perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S})$, then again (1) $O_i \not\perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S}, \boldsymbol{\lambda}_{\mathcal{I}_{O_i O_j|\boldsymbol{W},\boldsymbol{S}}})$ or (2) $\boldsymbol{\lambda}_{\mathcal{I}_{O_i|\boldsymbol{W},\boldsymbol{S}}} \not\perp\!\!\!\perp \boldsymbol{\lambda}_{\mathcal{I}_{O_j|\boldsymbol{W},\boldsymbol{S}}}$ (or both) holds by Proposition 3. Here, at least one member of $\{f(O_i | \boldsymbol{W}, \boldsymbol{S}), f(O_j | \boldsymbol{W}, \boldsymbol{S})\}$ is stationary, so $\boldsymbol{\lambda}_{\mathcal{I}_{O_i|\boldsymbol{W},\boldsymbol{S}}}$ or $\boldsymbol{\lambda}_{\mathcal{I}_{O_j|\boldsymbol{W},\boldsymbol{S}}}$ (or both) is a constant; hence, $\boldsymbol{\lambda}_{\mathcal{I}_{O_i|\boldsymbol{W},\boldsymbol{S}}} \perp\!\!\!\perp \boldsymbol{\lambda}_{\mathcal{I}_{O_j|\boldsymbol{W},\boldsymbol{S}}}$. Thus (1) must hold. We conclude that $O_i \not\perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S}, \boldsymbol{\lambda}_{\mathcal{I}_{O_i O_j|\boldsymbol{W},\boldsymbol{S}}})$, so $O_i^t \not\perp\!\!\!\perp O_j^t | (\boldsymbol{W}^t, \boldsymbol{S}^t)$ at some $t \in S(f(T))$. □

4.6.3 Mixture Faithfulness

Note that the reverse direction of Proposition 3 is not necessarily true; we do not necessarily have $O_i \perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S}, \boldsymbol{\lambda}_{\mathcal{I}_{O_i O_j|\boldsymbol{W},\boldsymbol{S}}})$ and $\boldsymbol{\lambda}_{\mathcal{I}_{O_i|\boldsymbol{W},\boldsymbol{S}}} \perp\!\!\!\perp \boldsymbol{\lambda}_{\mathcal{I}_{O_j|\boldsymbol{W},\boldsymbol{S}}}$, if $O_i \perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S})$. I however

believe that violating the reverse direction is extremely unlikely in practice. For example, take four binary variables $O_i, O_j, \boldsymbol{\lambda}_{\mathcal{I}_{O_i}}, \boldsymbol{\lambda}_{\mathcal{I}_{O_j}} \in \{0,1\}$ and one quaternary variable $\boldsymbol{\lambda}_{\mathcal{I}_{O_iO_j}} \in \{(0,0),(0,1),(1,0),(1,1)\}$. Let $\mathbb{P}_{\boldsymbol{\lambda}_{\mathcal{I}_{O_iO_j}}} = 0.25$, $\mathbb{P}_{\boldsymbol{\lambda}_{\mathcal{I}_{O_i}}} = 0.5$ and $\mathbb{P}_{\boldsymbol{\lambda}_{\mathcal{I}_{O_j}}} = 0.5$, so that we have $\boldsymbol{\lambda}_{\mathcal{I}_{O_i}} \perp\!\!\!\perp \boldsymbol{\lambda}_{\mathcal{I}_{O_j}}$ because $\mathbb{P}_{\boldsymbol{\lambda}_{\mathcal{I}_{O_iO_j}}} = \mathbb{P}_{\boldsymbol{\lambda}_{\mathcal{I}_{O_i}}\boldsymbol{\lambda}_{\mathcal{I}_{O_j}}} = \mathbb{P}_{\boldsymbol{\lambda}_{\mathcal{I}_{O_i}}}\mathbb{P}_{\boldsymbol{\lambda}_{\mathcal{I}_{O_j}}}$. Also consider the four probability tables in Table 4.2. Here, I have chosen the probabilities in the tables carefully by satisfying the following equation:

$$
\begin{aligned}
& f_m(O_i, O_j) = f_m(O_i)_m f(O_j) \\
\iff & \int f(O_i, O_j|\boldsymbol{\lambda}_{\mathcal{I}_{O_iO_j}})\; d\mathbb{P}_{\boldsymbol{\lambda}_{\mathcal{I}_{O_iO_j}}} = \\
& \qquad \int f(O_i|\boldsymbol{\lambda}_{\mathcal{I}_{O_i}})\; d\mathbb{P}_{\boldsymbol{\lambda}_{\mathcal{I}_{O_i}}} \int f(O_j|\boldsymbol{\lambda}_{\mathcal{I}_{O_j}})\; d\mathbb{P}_{\boldsymbol{\lambda}_{\mathcal{I}_{O_j}}} \\
\iff & 0.25\big[f(O_i, O_j|\boldsymbol{\lambda}_{\mathcal{I}_{O_iO_j}} = (0,0)) + f(O_i, O_j|\boldsymbol{\lambda}_{\mathcal{I}_{O_iO_j}} = (0,1)) + \\
& \qquad f(O_i, O_j|\boldsymbol{\lambda}_{\mathcal{I}_{O_iO_j}} = (1,0)) + f(O_i, O_j|\boldsymbol{\lambda}_{\mathcal{I}_{O_iO_j}} = (1,1))\big] = \\
& 0.25\big[f(O_i|\boldsymbol{\lambda}_{\mathcal{I}_{O_i}} = 0) f(O_j|\boldsymbol{\lambda}_{\mathcal{I}_{O_j}} = 0) + f(O_i|\boldsymbol{\lambda}_{\mathcal{I}_{O_i}} = 0) f(O_j|\boldsymbol{\lambda}_{\mathcal{I}_{O_j}} = 1) + \\
& \qquad f(O_i|\boldsymbol{\lambda}_{\mathcal{I}_{O_i}} = 1) f(O_j|\boldsymbol{\lambda}_{\mathcal{I}_{O_j}} = 0) + f(O_i|\boldsymbol{\lambda}_{\mathcal{I}_{O_i}} = 1) f(O_j|\boldsymbol{\lambda}_{\mathcal{I}_{O_j}} = 1)\big].
\end{aligned}
\tag{4.7}
$$

Of course, the above equality holds when we have conditional independence $f(O_i, O_j|\boldsymbol{\lambda}_{\mathcal{I}_{O_iO_j}} = (x,y)) = f(O_i|\boldsymbol{\lambda}_{\mathcal{I}_{O_i}} = x)\; f(O_j|\boldsymbol{\lambda}_{\mathcal{I}_{O_j}} = y), \forall x, y \in \{0,1\}$. We are however interested in the case when conditional dependence holds. We therefore instantiated the values of Tables 4.2a, 4.2b as well as the first three columns in Table 4.2c (the columns labeled (0,0), (0,1), (1,0)) such that $f(O_i, O_j|\boldsymbol{\lambda}_{\mathcal{I}_{O_iO_j}}) \neq f(O_i|\boldsymbol{\lambda}_{\mathcal{I}_{O_i}})\; f(O_j|\boldsymbol{\lambda}_{\mathcal{I}_{O_j}})$. We then solved for the fourth column using Equation 4.7 in order to complete Table 4.2c.

Notice that we obtain a unique value for the last column of Table 4.2c by solving Equation 4.7. Hence, each value in the last column of Table 4.2c has Lebesgue measure zero on the interval $[0,1]$, once we have defined all of the other values. Thus, $O_i \perp\!\!\!\perp O_j$ does not always imply that $O_i \perp\!\!\!\perp O_j|\boldsymbol{\lambda}_{\mathcal{I}_{O_iO_j}}$ and $\boldsymbol{\lambda}_{\mathcal{I}_{O_i}} \perp\!\!\!\perp \boldsymbol{\lambda}_{\mathcal{I}_{O_j}}$, but satisfying Equation 4.7 requires a very particular setup which is probably rarely encountered in practice.

Recall that we can construct a Lebesgue measure zero argument in general, when we have an algebraic equality such as $f_m(O_i, O_j|\boldsymbol{W}, \boldsymbol{S}) = f_m(O_i|\boldsymbol{W}, \boldsymbol{S}) f_m(O_j|\boldsymbol{W}, \boldsymbol{S})$ [Uhler et al., 2013]. This fact motivates the *mixture faithfulness* assumption:

	$\boldsymbol{\lambda}_{\mathcal{I}_{O_i}}$	
	0	1
$O_i = 0$	0.5	0.3
$O_i = 1$	0.5	0.7

(a)

	$\boldsymbol{\lambda}_{\mathcal{I}_{O_j}}$	
	0	1
$O_j = 0$	0.3	0.4
$O_j = 1$	0.7	0.6

(b)

	$\boldsymbol{\lambda}_{\mathcal{I}_{O_iO_j}}$			
	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$O_i = 0, O_j = 0$	0.2	0.25	0.04	0.07
$O_i = 0, O_j = 1$	0.3	0.25	0.26	0.23
$O_i = 1, O_j = 0$	0.1	0.25	0.16	0.33
$O_i = 1, O_j = 1$	0.4	0.25	0.54	0.37

(c)

Table 4.2: Example of a situation where $\boldsymbol{\lambda}_{\mathcal{I}_{O_i}} \perp\!\!\!\perp \boldsymbol{\lambda}_{\mathcal{I}_{O_j}}$ but $O_i \not\perp\!\!\!\perp O_j | \boldsymbol{\lambda}_{\mathcal{I}_{O_iO_j}}$ when $O_i \perp\!\!\!\perp O_j$. Table entries denote values of (a) $f(O_i|\boldsymbol{\lambda}_{\mathcal{I}_{O_i}})$, (b) $f(O_j|\boldsymbol{\lambda}_{\mathcal{I}_{O_j}})$ and (c) $f(O_i, O_j|\boldsymbol{\lambda}_{\mathcal{I}_{O_iO_j}})$.

Assumption 1. *A distribution $\mathbb{P}_{\boldsymbol{X}}$ is mixture faithful when the following property holds for any O_i, O_j and $\boldsymbol{W} \subseteq \boldsymbol{O} \setminus \{O_i, O_j\}$: if $O_i \perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S})$, then $O_i \perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S}, \boldsymbol{\lambda}_{\mathcal{I}_{O_i O_j | \boldsymbol{W}, \boldsymbol{S}}})$ and $\boldsymbol{\lambda}_{\mathcal{I}_{O_i | \boldsymbol{W}, \boldsymbol{S}}} \perp\!\!\!\perp \boldsymbol{\lambda}_{\mathcal{I}_{O_j | \boldsymbol{W}, \boldsymbol{S}}}$.*

Mixture faithfulness therefore corresponds to the reverse direction of Proposition 3 similar to how d-separation faithfulness corresponds to the reverse direction of the global directed Markov property. In other words:

Proposition 4. *Assume mixture faithfulness. Then, for any O_i, O_j and $\boldsymbol{W} \subseteq \boldsymbol{O} \setminus \{O_i, O_j\}$, we have $O_i \perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S})$ if and only if $O_i \perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S}, \boldsymbol{\lambda}_{\mathcal{I}_{O_i O_j | \boldsymbol{W}, \boldsymbol{S}}})$ and $\boldsymbol{\lambda}_{\mathcal{I}_{O_i | \boldsymbol{W}, \boldsymbol{S}}} \perp\!\!\!\perp \boldsymbol{\lambda}_{\mathcal{I}_{O_j | \boldsymbol{W}, \boldsymbol{S}}}$.*

We will find the mixture faithfulness assumption useful for the development of a new causal discovery algorithm under non-stationary feedback.

4.6.4 Conditional Independence Across Time

The mixture faithfulness assumption allows us to construct an important argument regarding conditional independence as a function of time when conditional independence holds in the mixture density. We have the following claim:

Lemma 4. *Assume mixture faithfulness. If $O_i \perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S})$, then $O_i^t \perp\!\!\!\perp O_j^t | (\boldsymbol{W}^t, \boldsymbol{S}^t)$ at all $t \in S(f(T))$.*

Proof. If $O_i \perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S})$, then $O_i \perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S}, \boldsymbol{\lambda}_{\mathcal{I}_{O_i O_j | \boldsymbol{W}, \boldsymbol{S}}})$ by mixture faithfulness. We therefore have $O_i \perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S}, \boldsymbol{\lambda}_{\mathcal{I}_{O_i O_j | \boldsymbol{W}, \boldsymbol{S}}} = c)$ for any c; in other words, $O_i^t \perp\!\!\!\perp O_j^t | (\boldsymbol{W}^t, \boldsymbol{S}^t)$ at any $t \in S(f(T))$. □

5.0 THE F^{2}CI ALGORITHM

I now use the ideas espoused in the previous chapter to develop a sound algorithm for causal discovery with mixture data. I believe that the proposed framework presents additional difficulties for causal discovery algorithms but a well-developed algorithm under this framework may help generalize causal discovery to scenarios involving non-stationary distributions and/or feedback loops.

5.1 POSSIBLE STRATEGIES

Any existing constraint-based causal discovery algorithm will unfortunately encounter substantial difficulty in performing causal discovery with mixture data, since conditional dependencies may arise either as a consequence of d-connection or parameter dependence (or both) rather than d-connection alone according to Lemma 1. We can nonetheless consider several alternative strategies for solving the causal discovery problem. I have identified three as described below.

One possible strategy for dealing with mixture data involves forgoing edge orientation altogether and focusing on skeleton and ancestor discovery which we can infer from conditional independence results alone. The detection of colliders naturally involves checking for conditional dependence which may be induced by parameter dependence rather than d-connection relations. We can however identify d-separation relations via mixture faithfulness (Lemma 4), so we may consider inferring graphical structure using only conditional independence as one possible strategy.

A second strategy involves relaxing the graph edge interpretations. For example, one may

consider modifying the interpretation of the arrowhead; specifically, if an edge between O_i and O_j has an arrowhead at O_j, then (1) $O_j^t \notin \boldsymbol{An}(O_i^t \cup \boldsymbol{S})$ at some $t \in S(f(T))$ or (2) there exists a dependency induced by parameter dependence (or both). I unfortunately find this strategy unsatisfying, since the connection between edge endpoints and the underlying causal graph becomes increasingly unclear if we use orientation rules to propagate the conditional dependence results.

A third strategy involves checking additional criteria in order to parse out the condition causing the conditional dependence. One such criterion involves checking the number of components in finite conditional mixture models, where I model finite mixtures of conditional densities and use the BIC score (or some similar score) to detect the number of components in each model. For example, suppose that $O_i \not\perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S})$. Then, I can check whether $f(O_i, O_j | \boldsymbol{W}, \boldsymbol{S})$ is stationary; i.e. whether the conditional density admits one component using say a mixture of linear regressions in the linear Gaussian CMJ case. If so, the conditional dependence must result from a d-connection rather than non-stationarity via Lemma 2. I prefer this last strategy and therefore will choose to pursue it further in this thesis.

5.2 THE MIXTURE APPROACH

I now provide additional details regarding the mixture modeling approach introduced in the last paragraph of the previous section. I believe that we can use a modified version of FCI along with conditional mixture modeling (e.g., via finite mixtures of linear regressions) to soundly infer causal structure between the variables in $\boldsymbol{O}$.

5.2.1 Endpoint Symbols

I will use seven different symbols at the edge endpoints throughout this thesis:

1. If we have the *unfilled arrowhead* $O_i * \!\!\rightarrow O_j$, then $O_j^t \notin \boldsymbol{An}(O_i^t \cup \boldsymbol{S}^t)$ for all time points $t \in S(f(T))$.
2. If we have the *unfilled tail* $O_i * \!\!- O_j$, then $O_j^t \in \boldsymbol{An}(O_i^t \cup \boldsymbol{S}^t)$ for all $t \in S(f(T))$.

3. If we have the *square* $O_i *\!\!-\!\!\blacksquare O_j$, then $O_j^t \in \boldsymbol{An}(O_i^t \cup \boldsymbol{S}^t)$ for time points $t \in A \subset S(f(T))$ and $O_j^t \notin \boldsymbol{An}(O_i^t \cup \boldsymbol{S}^t)$ for the other time points $t \in S(f(T)) \setminus A$.
4. If we have the *filled arrowhead* $O_i *\!\!-\!\!\blacktriangleright O_j$, then we either have $O_i *\!\!\rightarrow O_j$ or $O_i *\!\!-\!\!\blacksquare O_j$.
5. If we have the *filled tail* $O_i *\!\!-\!\!\bullet O_j$, then we either have $O_i *\!\!- O_j$ or $O_i *\!\!-\!\!\blacksquare O_j$.
6. If we have the *circle* $O_i *\!\!-\!\!\circ O_j$, then the endpoint at O_j has yet to be defined by the algorithm.
7. Finally, the *asterisk* $*$ denotes an unspecified endpoint, where we could have any one of the above 6 types of endpoints.

The proposed algorithm will output edges with endpoint symbols 1-6. Note that symbols 1, 2 and 6 also exist in the output of the original FCI algorithm. I introduce symbols 3, 4 and 5 when we work under the more general CMJ framework, so that the proposed algorithm may discover ancestral or non-ancestral relations across subsets of $S(f(T))$.

5.2.2 Skeleton Discovery

The first stage of constraint-based causal discovery involves skeleton discovery. The F^{2}CI algorithm will use FCI's skeleton discovery procedure on $\mathbb{P}_{\boldsymbol{X}^m}$. I therefore analyze the properties of FCI's skeleton discovery procedure when using $\mathbb{P}_{\boldsymbol{X}^m}$.

Denote the possible d-separating set of O_i at $t \in S(f(T))$ as $\boldsymbol{PDS}(O_i^t)$. We have $O_j^t \in \boldsymbol{PDS}(O_i^t)$ if and only if there exists a path π^t between O_i^t and O_j^t such that, for every subpath $\langle O_m^t, O_l^t, O_h^t \rangle$ of π^t, O_l^t is a collider on the subpath or $\langle O_m^t, O_l^t, O_h^t \rangle$ is a triangle. Next, denote the possible d-separating set of O_i as $\boldsymbol{PDS}(O_i)$, where $O_j \in \boldsymbol{PDS}(O_i)$ if and only if $O_j^t \in \boldsymbol{PDS}(O_i^t)$ at some $t \in S(f(T))$.

Now recall that PC's skeleton discovery and v-structure orientation procedures discover a superset of $\boldsymbol{PDS}(O_i^t), \forall i$ using $\mathbb{P}_{\boldsymbol{X}^t}$ under d-separation faithfulness with respect to $\mathbb{P}_{\boldsymbol{X}^t}$. I want to show that PC's skeleton discovery and v-structure orientation procedures will also discover a superset of $\boldsymbol{PDS}(O_i), \forall i$ using $\mathbb{P}_{\boldsymbol{X}^m}$ under d-separation faithfulness with respect to $\mathbb{P}_{\boldsymbol{X}^t}, \forall t \in S(f(T))$.

Consider the graph $\widetilde{\mathbb{G}}$ constructed by running PC's skeleton discovery and v-structure orientation procedures over $\boldsymbol{O}$ with a CI oracle. Let $O_k \in \widetilde{\boldsymbol{PDS}}(O_i)$ if and only if there

exists a path $\widetilde{\pi}$ in $\widetilde{\mathbb{G}}$ between O_i and O_k such that, for every subpath $\langle O_m, O_l, O_h \rangle$ of $\widetilde{\pi}$, O_l is a collider on the subpath or $\langle O_m, O_l, O_h \rangle$ is a triangle according to PC's skeleton discovery and v-structure orientation procedures. The following lemma shows that we can compute a superset of $\boldsymbol{PDS}(O_i), \forall i$ using PC's skeleton discovery and v-structure orientation procedures:

Lemma 5. *Under mixture faithfulness and d-separation faithfulness with respect to* $\mathbb{P}_{\boldsymbol{X}^t}, \forall t \in S(f(T))$*, PC's skeleton discovery and v-structure orientation procedures will discover a superset of* $\boldsymbol{PDS}(O_i), \forall i$ *with a CI oracle; i.e.,* $\boldsymbol{PDS}(O_i) \subseteq \widetilde{\boldsymbol{PDS}}(O_i), \forall i$.

Proof. I need to show that, for any arbitrary $O_k \in \boldsymbol{PDS}(O_i)$, we also have $O_k \in \widetilde{\boldsymbol{PDS}}(O_i)$. This means that whenever $\exists t \in S(f(T))$ with a path π_k^t between O_i^t and O_k^t such that, for every subpath $\langle O_m^t, O_l^t, O_h^t \rangle$ of π^t, O_l^t is a collider on the subpath or $\langle O_m^t, O_l^t, O_h^t \rangle$ is a triangle, then there also exists a path $\widetilde{\pi}_k$ in $\widetilde{\mathbb{G}}$ between O_i and O_k such that O_l is a collider on $\langle O_m, O_l, O_h \rangle$ or $\langle O_m, O_l, O_h \rangle$ is a triangle.

I first show that all adjacencies on π_k^t are on $\widetilde{\pi}_k$. Suppose not. Let $O_p^t \ast\!\!-\!\!\ast O_q^t$ denote an arbitrary adjacency present in π_k^t such that O_p and O_q are not adjacent in $\widetilde{\pi}_k$. Then, there exists a conditional independence $O_p \perp\!\!\!\perp O_q | (\boldsymbol{W}, \boldsymbol{S})$ when $O_p^t \not\perp\!\!\!\perp O_q^t | (\boldsymbol{W}^t, \boldsymbol{S}^t)$ by d-separation faithfulness. But this violates the mixture faithfulness assumption via the contrapositive of Lemma 4. Hence, every adjacency on π_k^t is also on $\widetilde{\pi}_k$.

I next show that if O_l^t is a collider on the subpath $\langle O_m^t, O_l^t, O_h^t \rangle$ on π_k^t, then O_l is either (1) a collider on the unshielded subpath $\langle O_m, O_l, O_h \rangle$ or (2) $\langle O_m, O_l, O_h \rangle$ is a triangle. Suppose first that O_l is on the unshielded subpath, so that O_m and O_h are non-adjacent in $\widetilde{\mathbb{G}}$; in other words, $O_m \perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S})$ for some $\boldsymbol{W} \subseteq \boldsymbol{O} \setminus \{O_i, O_j\}$. We already know that O_m and O_l as well as O_l and O_h are adjacent in $\widetilde{\mathbb{G}}$ by the previous paragraph. Now assume that $O_l \in \boldsymbol{W}$ for a contradiction. If O_l^t is a collider on the subpath $\langle O_m^t, O_l^t, O_h^t \rangle$ on π_k^t, then we may invoke Lemma 2.4 in [Colombo et al., 2012] to conclude that there exists an inducing path between O_m^t and O_l^t that is into O_l^t and there exists an inducing path between O_h^t and O_l^t that is into O_l^t. We may now invoke Lemma 2.5 in [Colombo et al., 2012] with the two inducing paths to conclude that O_m^t and O_h^t must be d-connected given $\{\boldsymbol{W}^t, \boldsymbol{S}^t\}$, so we have $O_m^t \not\perp\!\!\!\perp O_h^t | (\boldsymbol{W}^t, \boldsymbol{S}^t)$ by d-separation faithfulness. We finally conclude that $O_m \not\perp\!\!\!\perp O_h | (\boldsymbol{W}, \boldsymbol{S})$

by applying the contrapositive of Lemma 4; we have arrived at a contradiction. We therefore conclude that $O_l \notin \boldsymbol{W}$, so O_l must be a collider on the unshielded subpath in $\widetilde{\mathbb{G}}$. Finally, suppose that O_l is on the shielded subpath, so that O_m and O_h are adjacent in $\widetilde{\mathbb{G}}$. Then (2) holds because all three vertices in $\langle O_m, O_l, O_h \rangle$ must be adjacent again by the previous paragraph.

□

Recall that $\boldsymbol{PDS}(O_i^t)$ and $\boldsymbol{PDS}(O_j^t)$ are sufficiently large sets for determining the existence of an inducing path between O_i^t and O_j^t. Thus, we may claim that $\widetilde{\boldsymbol{PDS}}(O_i)$ and $\widetilde{\boldsymbol{PDS}}(O_j)$ are also sufficiently large for determining the existence of an inducing path between O_i^t and O_j^t for any $t \in S(f(T))$ by invoking Lemma 5.

We may therefore consider the following modified edge interpretations for the skeleton:

- The presence of an edge between two vertices O_i and O_j implies that $O_i \not\perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S})$ for all $\boldsymbol{W} \subseteq \widetilde{\boldsymbol{PDS}}(O_i)$ and for all $\boldsymbol{W} \subseteq \widetilde{\boldsymbol{PDS}}(O_j)$. Thus, there *may* exist an inducing path between O_i^t and O_j^t at some $t \in S(f(T))$ in this case by Lemma 1.
- The absence of an edge between two vertices O_i and O_j implies that $O_i \perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S})$ for some $\boldsymbol{W} \subseteq \boldsymbol{O} \setminus \{O_i, O_j\}$. Thus, there *cannot* exist an inducing path between O_i^t and O_j^t at any $t \in S(f(T))$ in this case by Lemma 4.

In other words, I can use FCI's skeleton discovery procedure equipped with a CI oracle over $\mathbb{P}_{\boldsymbol{X}^m}$ to rule out the existence of inducing paths.

5.2.3 V-Structure Discovery

Performing v-structure discovery will require the use of a conditional mixture modeling (CMM) oracle. Here, the CMM oracle outputs the number of components when given an arbitrary unconditional or conditional density. For example, if $f_m(O_i, O_j | \boldsymbol{W}, \boldsymbol{S}) = \frac{1}{3} f_1(O_i, O_j | \boldsymbol{W}, \boldsymbol{S}) + \frac{2}{3} f_2(O_i, O_j | \boldsymbol{W}, \boldsymbol{S})$ with $f_1(O_i, O_j | \boldsymbol{W}, \boldsymbol{S}) \neq f_2(O_i, O_j | \boldsymbol{W}, \boldsymbol{S})$, then the CMM oracle will output 2. Note that the CMM oracle always provides perfect answers just like the CI oracle.

We will find the CMM oracle useful because the oracle will allow us to differentiate between stationary and non-stationary densities. Let $f(\boldsymbol{A}|\boldsymbol{W},\boldsymbol{S})$ denote an arbitrary conditional density. If the conditional density $f(\boldsymbol{A}|\boldsymbol{W},\boldsymbol{S})$ is stationary over $S(f(T))$, then $f_m(\boldsymbol{A}|\boldsymbol{W},\boldsymbol{S})$ is a mixture of one conditional density from the family of conditional densities $\{f(\boldsymbol{A}^t|\boldsymbol{W}^t,\boldsymbol{S}^t) : t \in S(f(T))\}$, so the CMM oracle queried with $f_m(\boldsymbol{A}|\boldsymbol{W},\boldsymbol{S})$ will output one. On the other hand, if the conditional density $f(\boldsymbol{A}|\boldsymbol{W},\boldsymbol{S})$ is non-stationary, then $f_m(\boldsymbol{A}|\boldsymbol{W},\boldsymbol{S})$ is a mixture of more than one conditional density from $\{f(\boldsymbol{A}^t|\boldsymbol{W}^t,\boldsymbol{S}^t) \mid t \in S(f(T))\}$, so a CMM oracle queried with $f_m(\boldsymbol{A}|\boldsymbol{W},\boldsymbol{S})$ will output an integer greater than one.

We now have the following lemma for v-structure discovery which utilizes the concept of stationarity:

Lemma 6. *Assume d-separation faithfulness with respect to $\mathbb{P}_{\boldsymbol{X}^t}, \forall t \in S(f(T))$ and mixture faithfulness. Further assume that (1) $O_i \perp\!\!\!\perp O_j|(\boldsymbol{W},\boldsymbol{S})$ with $O_k \notin \boldsymbol{W}$, and (2) $O_i \not\perp\!\!\!\perp O_k|\boldsymbol{A}$ and $O_j \not\perp\!\!\!\perp O_k|\boldsymbol{A}$ where $\boldsymbol{A} = \{\boldsymbol{W} \setminus O_k\} \cup \boldsymbol{S}$. If both $f(O_i, O_k|\boldsymbol{A})$ and $f(O_j, O_k|\boldsymbol{A})$ are stationary, then $O_k^t \notin \boldsymbol{An}(\{O_i^t, O_j^t\} \cup \boldsymbol{S}^t)$ for all time points $t \in S(f(T))$. On the other hand, if (1) only $f(O_i, O_k|\boldsymbol{A})$ is stationary and at least one member of $\{f(O_k|\boldsymbol{A}), f(O_j|\boldsymbol{A})\}$ is stationary, or (2) only $f(O_j, O_k|\boldsymbol{A})$ is stationary and at least one member of $\{f(O_k|\boldsymbol{A}), f(O_i|\boldsymbol{A})\}$ is stationary, then $O_k^t \notin \boldsymbol{An}(\{O_i^t, O_j^t\} \cup \boldsymbol{S}^t)$ for some time point $t \in S(f(T))$.*

Proof. First assume that both $f(O_i, O_k|\boldsymbol{A})$ and $f(O_j, O_k|\boldsymbol{A})$ are stationary. If $O_i \perp\!\!\!\perp O_j|(\boldsymbol{W},\boldsymbol{S})$ with $O_k \notin \boldsymbol{W}$, then $O_i^t \perp\!\!\!\perp O_j^t|(\boldsymbol{W}^t,\boldsymbol{S}^t)$ and $O_k^t \notin \boldsymbol{W}^t$ at all $t \in S(f(T))$ by Lemma 4. Further, if (1) $O_i \not\perp\!\!\!\perp O_k|\boldsymbol{A}$, (2) $O_j \not\perp\!\!\!\perp O_k|\boldsymbol{A}$, (3) $f(O_i, O_k|\boldsymbol{A})$ is stationary and (4) $f(O_j, O_k|\boldsymbol{A})$ is stationary, then $O_i^t \not\perp\!\!\!\perp O_k^t|\boldsymbol{A}^t$ and $O_j^t \not\perp\!\!\!\perp O_k^t|\boldsymbol{A}^t$ at all $t \in S(f(T))$ by Lemma 2. Thus, at all $t \in S(f(T))$, we have (1) $O_i^t \perp\!\!\!\perp O_j^t|(\boldsymbol{W}^t,\boldsymbol{S}^t)$ with $O_k^t \notin \boldsymbol{W}^t$, and (2) $O_i^t \not\perp\!\!\!\perp O_k^t|\boldsymbol{A}^t$ and $O_j^t \not\perp\!\!\!\perp O_k^t|\boldsymbol{A}^t$. Now take any arbitrary time point t on $S(f(T))$ and invoke the argument for the forward direction in Lemma 3.1 of [Colombo et al., 2012] (minimality of $\boldsymbol{W}$ is not required for the forward direction).

Next assume that either (1) only $f(O_i, O_k|\boldsymbol{A})$ is stationary and at least one member of $\{f(O_k|\boldsymbol{A}), f(O_j|\boldsymbol{A})\}$ is stationary, or (2) only $f(O_j, O_k|\boldsymbol{A})$ is stationary and at least one member of $\{f(O_k|\boldsymbol{A}), f(O_i|\boldsymbol{A})\}$ is stationary. In either case, both $O_i^t \not\perp\!\!\!\perp O_k^t|\boldsymbol{A}^t$ and $O_j^t \not\perp\!\!\!\perp O_k^t|\boldsymbol{A}^t$ hold at some $t \in D \subseteq S(f(T))$ by Lemmas 2 and 3. Recall also that $O_i^t \perp\!\!\!\perp$

$O_j^t|(\boldsymbol{W}^t, \boldsymbol{S}^t)$ at all $t \in S(f(T))$ by Lemma 4. We can therefore take any arbitrary time point t in D and again invoke the argument for the forward direction in Lemma 3.1 of [Colombo et al., 2012]. □

In other words, we perform conditional mixture modeling of $f(O_i, O_k|\boldsymbol{A})$ and $f(O_j, O_k|\boldsymbol{A})$. We can then determine whether $f(O_i, O_k|\boldsymbol{A})$ or $f(O_j, O_k|\boldsymbol{A})$ (or both) admits one component. If a density only admits one component, then we can conclude that the density is stationary and orient the triple $\langle O_i, O_k, O_j \rangle$ accordingly.

I can thus use the above lemma to detect v-structures:

1. VSa: If (1) $O_i \perp\!\!\!\perp O_j|(\boldsymbol{W}, \boldsymbol{S})$ with $O_k \notin \boldsymbol{W}$, (2) $O_i \not\perp\!\!\!\perp O_k|\boldsymbol{A}$ and $O_j \not\perp\!\!\!\perp O_k|\boldsymbol{A}$ where $\boldsymbol{A} = \{\boldsymbol{W} \setminus O_k\} \cup \boldsymbol{S}$, and (3) both $f(O_i, O_k|\boldsymbol{A})$ and $f(O_j, O_k|\boldsymbol{A})$ are stationary, then orient $O_i \ast\!\!-\!\!\circ O_k \circ\!\!-\!\!\ast O_j$, $O_i \ast\!\!-\!\!\blacktriangleright O_k \circ\!\!-\!\!\ast O_j$, $O_i \ast\!\!-\!\!\circ O_k \blacktriangleleft\!\!-\!\!\ast O_j$, $O_i \ast\!\!-\!\!\blacktriangleright O_k \blacktriangleleft\!\!-\!\!\ast O_j$, $O_i \ast\!\!\rightarrow O_k \blacktriangleleft\!\!-\!\!\ast O_j$ or $O_i \ast\!\!-\!\!\blacktriangleright O_k \leftarrow\!\!\ast O_j$ as $O_i \ast\!\!\rightarrow O_k \leftarrow\!\!\ast O_j$.
2. VSb: If (1) $O_i \perp\!\!\!\perp O_j|(\boldsymbol{W}, \boldsymbol{S})$ with $O_k \notin \boldsymbol{W}$, (2) $O_i \not\perp\!\!\!\perp O_k|\boldsymbol{A}$ and $O_j \not\perp\!\!\!\perp O_k|\boldsymbol{A}$ where $\boldsymbol{A} = \{\boldsymbol{W} \setminus O_k\} \cup \boldsymbol{S}$, and (3) either only $f(O_i, O_k|\boldsymbol{A})$ is stationary and at least one member of $\{f(O_k|\boldsymbol{A}), f(O_j|\boldsymbol{A})\}$ is stationary, or only $f(O_j, O_k|\boldsymbol{A})$ is stationary and at least one member of $\{f(O_k|\boldsymbol{A}), f(O_i|\boldsymbol{A})\}$ is stationary, then orient the triple $\langle O_i, O_k, O_j \rangle$ as either (1) $O_i \ast\!\!-\!\!\blacktriangleright O_k \leftarrow\!\!\ast O_j$ when $O_i \ast\!\!-\!\!\circ O_k \leftarrow\!\!\ast O_j$, (2) $O_i \ast\!\!\rightarrow O_k \blacktriangleleft\!\!-\!\!\ast O_j$ when $O_i \ast\!\!\rightarrow O_k \circ\!\!-\!\!\ast O_j$, (3) $O_i \ast\!\!-\!\!\blacktriangleright O_k \blacktriangleleft\!\!-\!\!\ast O_j$ when $O_i \ast\!\!-\!\!\circ O_k \circ\!\!-\!\!\ast O_j$, (4) $O_i \ast\!\!-\!\!\blacksquare O_k \leftarrow\!\!\ast O_j$ when $O_i \ast\!\!-\!\!\bullet O_k \leftarrow\!\!\ast O_j$, (5) $O_i \ast\!\!\rightarrow O_k \blacksquare\!\!-\!\!\ast O_j$ when $O_i \ast\!\!\rightarrow O_k \bullet\!\!-\!\!\ast O_j$ or (6) $O_i \ast\!\!-\!\!\blacksquare O_k \blacksquare\!\!-\!\!\ast O_j$ when $O_i \ast\!\!-\!\!\bullet O_k \bullet\!\!-\!\!\ast O_j$.

Notice that I enumerated all possible outputs in VSb. I now choose not to spell out all possibilities to keep the presentation short. However, we should keep the following statements in mind:

1. If we orient an endpoint on a pre-existent circle, then we obtain the endpoint. For example, if $O_i \ast\!\!-\!\!\circ O_j$ and the algorithm demands $O_i \ast\!\!\rightarrow O_j$, then we obtain $O_i \ast\!\!\rightarrow O_j$.
2. If we orient a filled tail or arrowhead on a pre-existent unfilled tail or arrowhead, respectively, then we always keep the unfilled counterpart. For example, if $O_i \ast\!\!-\!\!\blacktriangleright O_j$ and the algorithm demands $O_i \ast\!\!\rightarrow O_j$, then we obtain $O_i \ast\!\!\rightarrow O_j$. We cannot orient a filled tail on an unfilled arrowhead, or a filled arrowhead on an unfilled tail because these situations violate the edge interpretations.

3. If we orient a filled arrowhead on a pre-existent filled tail, then we obtain a square. Similarly, if we orient a filled tail on a pre-existent filled arrowhead, then we also obtain a square. For example, if $O_i \ast\!\!\blacktriangleright O_j$ and the algorithm demands $O_i \ast\!\!-\!\!\bullet O_j$, then we obtain $O_i \ast\!\!-\!\!\blacksquare O_j$.

I therefore write VSb more compactly as follows: if (1) $O_i \perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S})$ with $O_k \notin \boldsymbol{W}$, (2) $O_i \not\perp\!\!\!\perp O_k | \boldsymbol{A}$ and $O_j \not\perp\!\!\!\perp O_k | \boldsymbol{A}$ where $\boldsymbol{A} = \{\boldsymbol{W} \setminus O_k\} \cup \boldsymbol{S}$, and (3) either only $f(O_i, O_k|\boldsymbol{A})$ is stationary and at least one member of $\{f(O_k|\boldsymbol{A}), f(O_j|\boldsymbol{A})\}$ is stationary, or only $f(O_j, O_k|\boldsymbol{A})$ is stationary and at least one member of $\{f(O_k|\boldsymbol{A}), f(O_i|\boldsymbol{A})\}$ is stationary, then orient the triple $\langle O_i, O_k, O_j \rangle$ as $O_i \ast\!\!\blacktriangleright O_k \blacktriangleleft\!\!\ast\, O_j$.

5.2.4 Fourth Orientation Rule

I will now move on to proving the soundness of the proposed algorithm's orientation rules. In general, I adopt the following strategy for the orientation rules: for each rule in FCI, (1) eliminate all endpoints in the rule which are not required in the proofs of soundness in [Zhang, 2008], and then (2) combine the minimalist rule with all (logically useful) combinations of stationary and non-stationary densities.

Now orientation rules 1, 4, 5, 9 and 10 require longer arguments than the other rules, so I will present the arguments for rules 1, 4, 5, 9 and 10 in separate subsections. I will order the presentation of the rules according to a logical progression, rather then simply present the rules in numerical order. I will first cover rule 4 in this subsection, then rule 1 in Subsection 5.2.5, then rules 5, 9 and 10 in Subsection 5.2.6, and finally the remaining rules in Subsection 5.2.7.

Now the fourth orientation rule requires the following lemma:

Lemma 7. *Assume d-separation faithfulness with respect to* $\mathbb{P}_{\boldsymbol{X}^t}, \forall t \in S(f(T))$ *and mixture faithfulness. Let* $\pi_{ik} = \{O_i, \ldots, O_l, O_j, O_k\}$ *be a sequence of at least four vertices which may satisfy the following:*

A1. $O_i \perp\!\!\!\perp O_k | (\boldsymbol{W}, \boldsymbol{S})$ *with* $\boldsymbol{W} \subseteq \boldsymbol{O} \setminus \{O_i, O_k\}$,

A2. Any two successive vertices O_h *and* O_{h+1} *on* π_{ik} *are conditionally dependent given* $(\boldsymbol{Y} \setminus \{O_h, O_{h+1}\}) \cup \boldsymbol{S}$ *for all* $\boldsymbol{Y} \subseteq \boldsymbol{W}$,

A3. *We have the stationary density* $f(O_h, O_{h+1}|((\boldsymbol{Y} \setminus \{O_h, O_{h+1}\}) \cup \boldsymbol{S})$ *for all* $\boldsymbol{Y} \subseteq \boldsymbol{W}$ *for any two successive vertices* O_h *and* O_{h+1} *on* π_{ik},

A4. *All vertices* O_h *between* O_i *and* O_j *(not including* O_i *and* O_j*) satisfy* $O_h^t \in \boldsymbol{An}(O_k^t)$ *and* $O_h^t \notin \boldsymbol{An}(\{O_{h-1}^t, O_{h+1}^t\} \cup \boldsymbol{S}^t)$ *for all time points* $t \in S(f(T))$*, where* O_{h-1} *and* O_{h+1} *denote the vertices adjacent to* O_h *on* π_{ik}.

The following conclusions hold, ordered from strongest to weakest:

C1. *Assume A1-A4 hold exactly. If* $O_j \in \boldsymbol{W}$*, then (a1)* $O_j^t \in \boldsymbol{An}(O_k^t \cup \boldsymbol{S}^t)$ *at all* $t \in S(f(T))$ *and (a2)* $O_k^t \in \boldsymbol{An}(O_j^t \cup \boldsymbol{S}^t)$ *at all* $t \in S(f(T))$*. If* $O_j \notin \boldsymbol{W}$*, then (a3)* $O_j^t \notin \boldsymbol{An}(\{O_l^t, O_k^t\} \cup \boldsymbol{S}^t)$ *at all* $t \in S(f(T))$ *and (a4)* $O_k^t \notin \boldsymbol{An}(O_j^t \cup \boldsymbol{S}^t)$ *at all* $t \in S(f(T))$.

C2. *Assume A1-A4 hold except, for A3, suppose there exists one and only one non-stationary density* $f(O_h, O_{h+1}|\boldsymbol{Y} \setminus \{O_h, O_{h+1}\} \cup \boldsymbol{S})$ *for some* $\boldsymbol{Y} \subseteq \boldsymbol{W}$*. However, assume that at least one member of* $\{f(O_h|\boldsymbol{Y} \setminus \{O_h, O_{h+1}\} \cup \boldsymbol{S}),\ f(O_{h+1}|\boldsymbol{Y} \setminus \{O_h, O_{h+1}\} \cup \boldsymbol{S})\}$ *is stationary. If* $O_j \in \boldsymbol{W}$*, then (b1)* $O_j^t \in \boldsymbol{An}(O_k^t \cup \boldsymbol{S}^t)$ *at some* $t \in S(f(T))$ *and (b2)* $O_k^t \notin \boldsymbol{An}(O_j^t \cup \boldsymbol{S}^t)$ *at all* $t \in S(f(T))$*. If* $O_j \notin \boldsymbol{W}$*, then (b3)* $O_j^t \notin \boldsymbol{An}(\{O_l^t, O_k^t\} \cup \boldsymbol{S}^t)$ *at some* $t \in S(f(T))$ *and (b4)* $O_k^t \notin \boldsymbol{An}(O_j^t \cup \boldsymbol{S}^t)$ *at all* $t \in S(f(T))$.

C3. *Assume A1-A4 hold except, for A4, suppose* $\exists t' \in S(f(T))$ *where all vertices* O_h *between* O_i *and* O_j *(not including* O_i *and* O_j*) satisfy* $O_h^{t'} \in \boldsymbol{An}(O_k^{t'})$ *and* $O_h^{t'} \notin \boldsymbol{An}(\{O_{h-1}^{t'}, O_{h+1}^{t'}\} \cup \boldsymbol{S}^{t'})$*. If* $O_j \in \boldsymbol{W}$*, then (c1)* $O_j^t \in \boldsymbol{An}(O_k^t \cup \boldsymbol{S}^t)$ *at some* $t \in S(f(T))$ *and (c2)* $O_k^t \notin \boldsymbol{An}(O_j^t \cup \boldsymbol{S}^t)$ *at some* $t \in S(f(T))$*. If* $O_j \notin \boldsymbol{W}$*, then (c3)* $O_j^t \notin \boldsymbol{An}(\{O_l^t, O_k^t\} \cup \boldsymbol{S}^t)$ *at some* $t \in S(f(T))$ *and (c4)* $O_k^t \notin \boldsymbol{An}(O_j^t \cup \boldsymbol{S}^t)$ *at some* $t \in S(f(T))$.

Proof. We can claim that A1 and A2 hold at all $t \in S(f(T))$ by invoking Lemma 4 for A1 and Lemma 2 with A3 for A2. Observe that the conditions in A4 hold at all $t \in S(f(T))$. By acyclicity of the CMJ-DAG at any one time point, conclusions (a1)-(a4) follow due to Lemma 3.2 in [Colombo et al., 2012] applied at any arbitrary $t \in S(f(T))$.

The proof for conclusions (c1)-(c4) proceeds similarly except A4 only holds at some $t \in S(f(T))$. We can therefore invoke Lemma 3.2 in [Colombo et al., 2012] again but only applied at some $t \in S(f(T))$, where all of the sufficient conditions are satisfied.

Let us now tackle conclusions (b1) and (b3) (but not (b2) and (b4)). Notice that A2 only holds at some $t \in S(f(T))$ by Lemmas 2 and 3, but A1 and A4 hold at all $t \in S(f(T))$

in this case. We can therefore invoke Lemma 3.2 in [Colombo et al., 2012] but only applied at some $t \in S(f(T))$, where all of the sufficient conditions are satisfied.

We now prove conclusions (b2) and (b4). From A4, we know that $O_l^t \in \boldsymbol{An}(O_k^t)$ and $O_l^t \notin \boldsymbol{An}(\boldsymbol{S}^t)$ at all $t \in S(f(T))$. Hence, we must have $O_k^t \notin \boldsymbol{An}(\boldsymbol{S}^t)$ at all $t \in S(f(T))$. We therefore only need to prove that $O_k^t \notin \boldsymbol{An}(O_j^t)$ at all $t \in S(f(T))$. Assume contrary to the claim that $O_k^t \in \boldsymbol{An}(O_j^t)$ at some $t \in S(f(T))$. Recall that $O_l^t \in \boldsymbol{An}(O_k^t)$ at all $t \in S(f(T))$ from A4, so $O_l^t \in \boldsymbol{An}(O_j^t)$ also at some $t \in S(f(T))$. This statement however contradicts another part of A4, where we must have $O_l^t \notin \boldsymbol{An}(O_j^t)$ at all $t \in S(f(T))$. Hence $O_k^t \notin \boldsymbol{An}(O_j^t)$ at all $t \in S(f(T))$.

□

We can use the above lemma to apply the following three orientation rules:

1. R4a: Suppose (1) there exists a path $\pi_{ik} = \{O_i, \ldots, O_l, O_j, O_k\}$, (2) $O_i \perp\!\!\!\perp O_k | (\boldsymbol{W}, \boldsymbol{S})$, (3) any two successive vertices O_h and O_{h+1} on π_{ik} are conditionally dependent given $(\boldsymbol{Y} \setminus \{O_h, O_{h+1}\}) \cup \boldsymbol{S}$ for all $\boldsymbol{Y} \subseteq \boldsymbol{W}$, (4) we have the stationary density $f(O_h, O_{h+1} | ((\boldsymbol{Y} \setminus \{O_h, O_{h+1}\}) \cup \boldsymbol{S})$ for all $\boldsymbol{Y} \subseteq \boldsymbol{W}$ for any two successive vertices O_h and O_{h+1} on π_{ik}, (5) all vertices O_h between O_i and O_j (not including O_i and O_j) satisfy $O_h^t \in \boldsymbol{An}(O_k^t)$ and $O_h^t \notin \boldsymbol{An}(\{O_{h-1}^t, O_{h+1}^t\} \cup \boldsymbol{S}^t)$ for all time points $t \in S(f(T))$, where O_{h-1} and O_{h+1} denote the vertices adjacent to O_h on π_{ik}. If further $O_j \in \boldsymbol{W}$, then orient $O_j \circ\!\!-\!\!* O_k$ or $O_j \bullet\!\!-\!\!* O_k$ as $O_j \to O_k$. Otherwise, if $O_j \notin \boldsymbol{W}$, then orient the triple $\langle O_l, O_j, O_k \rangle$ as $O_l \leftrightarrow O_j \leftrightarrow O_k$.
2. R4b: Suppose (1)-(5) hold in R4a except, for (4), suppose there exists one and only one non-stationary density $f(O_h, O_{h+1} | (\boldsymbol{Y} \setminus \{O_h, O_{h+1}\}) \cup \boldsymbol{S})$ for some $\boldsymbol{Y} \subseteq \boldsymbol{W}$. However, assume that at least one member of $\{f(O_h | (\boldsymbol{Y} \setminus \{O_h, O_{h+1}\}) \cup \boldsymbol{S}),\ f(O_{h+1} | (\boldsymbol{Y} \setminus \{O_h, O_{h+1}\}) \cup \boldsymbol{S})\}$ is stationary. Further, if $O_j \in \boldsymbol{W}$, then orient $O_j \circ\!\!-\!\!* O_k$ or $O_j \leftarrow\!\!* O_k$ as $O_j \bullet\!\!\to O_k$. Otherwise, if $O_j \notin \boldsymbol{W}$, then orient the triple $\langle O_l, O_j, O_k \rangle$ as $O_l \leftrightarrow O_j \leftrightarrow O_k$.
3. R4c: Suppose (1)-(5) hold in R4a except, for (5), suppose $\exists t' \in S(f(T))$ where all vertices O_h between O_i and O_j (not including O_i and O_j) satisfy $O_h^{t'} \in \boldsymbol{An}(O_k^{t'})$ and $O_h^{t'} \notin \boldsymbol{An}(\{O_{h-1}^{t'}, O_{h+1}^{t'}\} \cup \boldsymbol{S}^{t'})$. Further, if $O_j \in \boldsymbol{W}$, then orient $O_j \circ\!\!-\!\!* O_k$ or $O_j \leftarrow\!\!* O_k$ as $O_j \bullet\!\!\to O_k$. Otherwise, if $O_j \notin \boldsymbol{W}$, then orient the triple $\langle O_l, O_j, O_k \rangle$ as $O_l *\!\!\to O_j \leftarrow\!\!\to O_k$.

Let us pay special attention to R4c. Here, we may identify the existence of the time point t' when there exists only one filled endpoint on the discriminating path or, more generally, when the discriminating path is *contemporaneous*:

Definition 5. *Consider a path $\pi = \langle O_1, \ldots, O_n \rangle$. Also consider a similar path $\mathcal{E} = \langle O_1, \ldots, O_n \rangle$ but with only unfilled endpoints. Here, an endpoint on π may have a corresponding unfilled endpoint in $\mathcal{E}$. I say that π is contemporaneous according to $\mathcal{E}$ if and only if $\exists t' \in S(f(T))$ such that, for every arbitrary endpoint on π with a corresponding unfilled endpoint in $\mathcal{E}$, say at O_i on the edge between O_i and O_j, we have (1) $O_i^{t'} \notin \boldsymbol{An}(O_j^{t'} \cup \boldsymbol{S}^{t'})$ when we have a corresponding unfilled arrowhead in $\mathcal{E}$, or (2) $O_i^{t'} \in \boldsymbol{An}(O_j^{t'} \cup \boldsymbol{S}^{t'})$ when we have a corresponding unfilled tail in $\mathcal{E}$.*

In other words, π is contemporaneous according to $\mathcal{E}$ when the unfilled endpoints in $\mathcal{E}$ correspond to endpoints on π with ancestral/non-ancestral relations that exist at the same point in time. For example, suppose we create the v-structure $O_i \ast\!\!\rightarrow O_j \leftarrow\!\!\ast\, O_k$ with VSb. We then have a contemporaneous path $O_i \ast\!\!\rightarrow O_j \leftarrow\!\!\ast\, O_k$ according to $O_i \ast\!\!\rightarrow O_j \leftarrow\!\!\ast O_k$. The set $\mathcal{E}$ therefore corresponds to an unshielded v-structure with unfilled endpoints in this case. Similarly, the set $\mathcal{E}$ corresponds to a discriminating path with unfilled endpoints in R4c.

5.2.5 First Orientation Rule

We require the definition of a minimal independence set for the first orientation rule:

Definition 6. *If $O_i \perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S})$ with $\boldsymbol{W} \subseteq \boldsymbol{O} \setminus \{O_i, O_j\}$, but we have $O_i \not\perp\!\!\!\perp O_j | (\boldsymbol{B}, \boldsymbol{S})$ for any $\boldsymbol{B} \subset \boldsymbol{W}$, then $\boldsymbol{W}$ is a minimal independence set for O_i and O_j.*

Recall that Proposition 3 tells us that we have $O_i \not\perp\!\!\!\perp O_j | (\boldsymbol{B}, \boldsymbol{S})$ in the following three cases:

List 1. *We have:*

1. $O_i^t \perp\!\!\!\perp O_j^t | (\boldsymbol{B}^t, \boldsymbol{S}^t)$ *at all* $t \in S(f(T))$ *with* $\boldsymbol{\lambda}_{\mathcal{I}_{O_i | \boldsymbol{B}, \boldsymbol{S}}} \not\perp\!\!\!\perp \boldsymbol{\lambda}_{\mathcal{I}_{O_j | \boldsymbol{B}, \boldsymbol{S}}}$
2. $O_i^t \not\perp\!\!\!\perp O_j^t | (\boldsymbol{B}^t, \boldsymbol{S}^t)$ *at some* $t \in S(f(T))$ *with* $\boldsymbol{\lambda}_{\mathcal{I}_{O_i | \boldsymbol{B}, \boldsymbol{S}}} \perp\!\!\!\perp \boldsymbol{\lambda}_{\mathcal{I}_{O_j | \boldsymbol{B}, \boldsymbol{S}}}$
3. $O_i^t \not\perp\!\!\!\perp O_j^t | (\boldsymbol{B}^t, \boldsymbol{S}^t)$ *at some* $t \in S(f(T))$ *with* $\boldsymbol{\lambda}_{\mathcal{I}_{O_i | \boldsymbol{B}, \boldsymbol{S}}} \not\perp\!\!\!\perp \boldsymbol{\lambda}_{\mathcal{I}_{O_j | \boldsymbol{B}, \boldsymbol{S}}}$

We would however like to claim that if $O_i \perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S})$ with $\boldsymbol{W}$ having minimal cardinality, then $O_i^t \not\perp\!\!\!\perp O_j^t | (\boldsymbol{B}^t, \boldsymbol{S}^t)$ at some $t \in S(f(T))$ for all $\boldsymbol{B} \subset \boldsymbol{W}$. However, the first item in the above list introduces a problem because we have $O_i^t \perp\!\!\!\perp O_j^t | (\boldsymbol{B}^t, \boldsymbol{S}^t)$ at all $t \in S(f(T))$.

We can nonetheless make progress by realizing that satisfying the first item in List 1 is very difficult in practice. Let $\boldsymbol{Z} = \boldsymbol{W} \setminus \boldsymbol{B}$. Consider the following definition:

Definition 7. *A non-empty variable set $\boldsymbol{Z} \subseteq \boldsymbol{W}$ is parameter independence inducing (PII) if and only if (1) $\boldsymbol{W}$ is the smallest set such that $O_i^t \perp\!\!\!\perp O_j^t | (\boldsymbol{W}^t, \boldsymbol{S}^t)$ at all $t \in S(f(T))$ with $\boldsymbol{\lambda}_{\mathcal{I}_{O_i|\boldsymbol{W},\boldsymbol{S}}} \perp\!\!\!\perp \boldsymbol{\lambda}_{\mathcal{I}_{O_j|\boldsymbol{W},\boldsymbol{S}}}$, but (2) $O_i^t \perp\!\!\!\perp O_j^t | (\boldsymbol{W}^t \setminus \boldsymbol{Z}^t, \boldsymbol{S}^t)$ at all $t \in S(f(T))$ with $\boldsymbol{\lambda}_{\mathcal{I}_{O_i|\boldsymbol{W}\setminus\boldsymbol{Z},\boldsymbol{S}}} \not\perp\!\!\!\perp \boldsymbol{\lambda}_{\mathcal{I}_{O_j|\boldsymbol{W}\setminus\boldsymbol{Z},\boldsymbol{S}}}$.*

In other words, if conditional independence holds given $\{\boldsymbol{W}^t \setminus \boldsymbol{Z}^t, \boldsymbol{S}^t\}$ for all $t \in S(f(T))$, then introducing $\boldsymbol{Z}$ maintains the conditional independence but also induces parameter independence.

Now I believe that PII sets almost never occur in practice. For example, consider the CMJ in Figure 5.1a. Assume that $f_t(O_i, O_j)$ uniquely changes at the dotted red lines in Figure 5.1a. Then, we have $O_1^t \perp\!\!\!\perp O_2^t$ at all $t \in S(f(T))$ by the global directed Markov property but parameter dependence may hold. We can however consider a variable set $\boldsymbol{Z}$ that takes four distinct values between the dotted red lines in the intervals $[0, 1)$, $[1, 2)$, $[2, 3)$ and $[3, 4)$. Then $O_1^t \perp\!\!\!\perp O_2^t | \boldsymbol{Z}^t$ at all $t \in S(f(T))$ and $\boldsymbol{\lambda}_{\mathcal{I}_{O_1|\boldsymbol{Z}}} \perp\!\!\!\perp \boldsymbol{\lambda}_{\mathcal{I}_{O_2|\boldsymbol{Z}}}$, so $\boldsymbol{Z}$ induces parameter independence. I can make a similar statement for another $\boldsymbol{Z}$ in Figure 5.1b that takes five values.

Notice that choosing $\boldsymbol{Z}$ depends on $\mathbb{P}_T$. If nature constructs a PII variable set, then the variable set must correspond with the investigator-determined $\mathbb{P}_T$. On the other hand, if an investigator constructs a PII variable set, then he or she must design $\mathbb{P}_T$ using prior knowledge about the underlying CMJ. I believe both of the aforementioned cases are uncommon. I therefore feel safe to make the following assumption:

Assumption 2. *Parameter faithfulness holds if and only if we cannot query the CI oracle with any PII variable set.*

Observe that we may also construct a Lebesgue measure zero argument for parameter faithfulness, as we did with mixture faithfulness, since we must have the algebraic equality

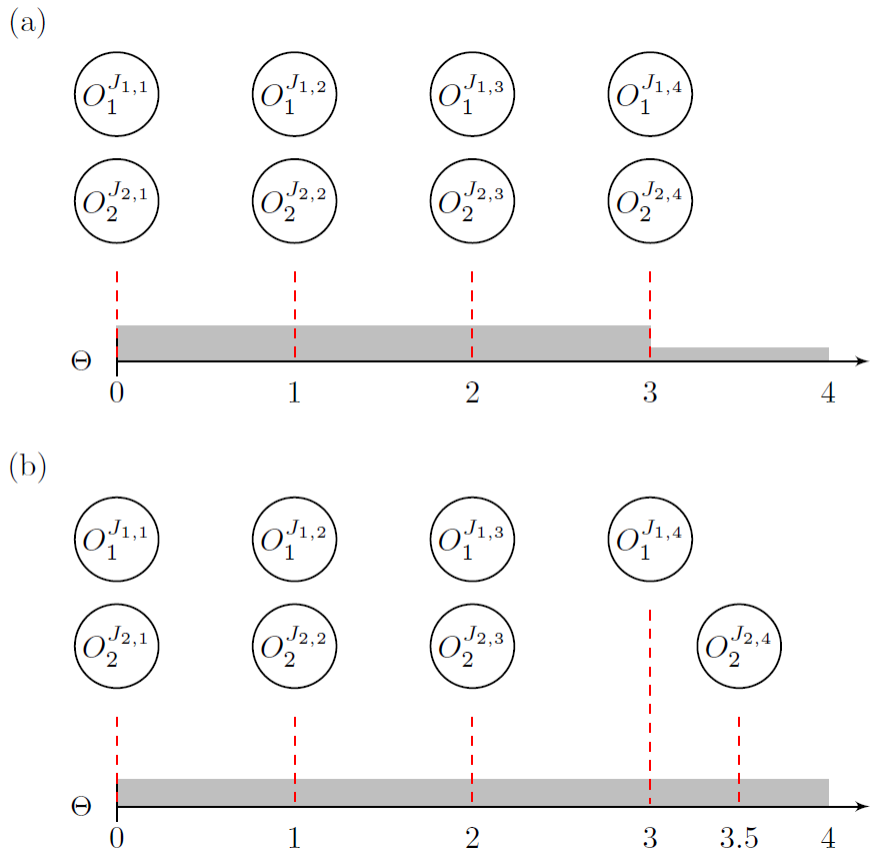


Figure 5.1: (a) An example of a CMJ and a time distribution $\mathbb{P}^T$. A variable taking on 4 unique values between the dotted red lines ensures that parameter independence holds. (b) We can likewise consider a variable taking 5 unique values in this shifted case.

$\mathbb{P}_{\boldsymbol{\lambda}_{\mathcal{I}_{O_i|\boldsymbol{W},\boldsymbol{S}}}\boldsymbol{\lambda}_{\mathcal{I}_{O_j|\boldsymbol{W},\boldsymbol{S}}}} = \mathbb{P}_{\boldsymbol{\lambda}_{\mathcal{I}_{O_i|\boldsymbol{W},\boldsymbol{S}}}}\mathbb{P}_{\boldsymbol{\lambda}_{\mathcal{I}_{O_j|\boldsymbol{W},\boldsymbol{S}}}}$ when parameter independence holds.

Parameter faithfulness has another close connection with mixture faithfulness. Mixture faithfulness allows us to claim the following if and only if statement by Proposition 4: we have $O_i \perp\!\!\!\perp O_j|(\boldsymbol{W},\boldsymbol{S})$ if and only if $O_i \perp\!\!\!\perp O_j|(\boldsymbol{W},\boldsymbol{S},\boldsymbol{\lambda}_{\mathcal{I}_{O_iO_j|\boldsymbol{W},\boldsymbol{S}}})$ and $\boldsymbol{\lambda}_{\mathcal{I}_{O_i|\boldsymbol{W},\boldsymbol{S}}} \perp\!\!\!\perp \boldsymbol{\lambda}_{\mathcal{I}_{O_j|\boldsymbol{W},\boldsymbol{S}}}$. We may re-write the above statement in terms of $\boldsymbol{B} \subset \boldsymbol{W}$ as follows: $O_i \not\perp\!\!\!\perp O_j|(\boldsymbol{B},\boldsymbol{S})$ if and only if $O_i^t \not\perp\!\!\!\perp O_j^t|(\boldsymbol{B}^t,\boldsymbol{S}^t)$ for some $t \in S(f(T))$ or $\boldsymbol{\lambda}_{\mathcal{I}_{O_i|\boldsymbol{B},\boldsymbol{S}}} \not\perp\!\!\!\perp \boldsymbol{\lambda}_{\mathcal{I}_{O_j|\boldsymbol{B},\boldsymbol{S}}}$. The "or" logical disjunction again presents the problem here because we may have $O_i^t \perp\!\!\!\perp O_j^t|(\boldsymbol{B}^t,\boldsymbol{S}^t)$ for all $t \in S(f(T))$ but $\boldsymbol{\lambda}_{\mathcal{I}_{O_i|\boldsymbol{B},\boldsymbol{S}}} \not\perp\!\!\!\perp \boldsymbol{\lambda}_{\mathcal{I}_{O_j|\boldsymbol{B},\boldsymbol{S}}}$. However, the parameter faithfulness assumption allows us to avoid this problematic case when used in conjunction with mixture faithfulness:

Proposition 5. *Assume that mixture faithfulness and parameter faithfulness holds. If* $O_i \perp\!\!\!\perp O_j|(\boldsymbol{W},\boldsymbol{S})$ *with minimal independence set* $\boldsymbol{W}$*, then* $\exists t \in S(f(T))$ *for each* $\boldsymbol{B} \subset \boldsymbol{W}$ *such that* $O_i^t \not\perp\!\!\!\perp_d O_j^t|(\boldsymbol{B}^t,\boldsymbol{S}^t)$*; note that* t *may not necessarily be the same for all* $\boldsymbol{B} \subset \boldsymbol{W}$*.*

Proof. If $O_i \perp\!\!\!\perp O_j|(\boldsymbol{W},\boldsymbol{S})$, then $O_i^t \perp\!\!\!\perp O_j^t|(\boldsymbol{W}^t,\boldsymbol{S}^t)$ at all $t \in S(f(T))$ with $\boldsymbol{\lambda}_{\mathcal{I}_{O_i|\boldsymbol{W},\boldsymbol{S}}} \perp\!\!\!\perp \boldsymbol{\lambda}_{\mathcal{I}_{O_j|\boldsymbol{W},\boldsymbol{S}}}$ by mixture faithfulness. By parameter faithfulness, if we remove any non-empty variable set $\boldsymbol{Z}$ from $\boldsymbol{W}$, then we cannot have $O_i^t \perp\!\!\!\perp O_j^t|(\boldsymbol{W}^t \setminus \boldsymbol{Z}^t,\boldsymbol{S}^t)$ at all $t \in S(f(T))$ with $\boldsymbol{\lambda}_{\mathcal{I}_{O_i|\boldsymbol{W},\boldsymbol{S}}} \not\perp\!\!\!\perp \boldsymbol{\lambda}_{\mathcal{I}_{O_j|\boldsymbol{W},\boldsymbol{S}}}$. Therefore, by minimality of $\boldsymbol{W}$ and Lemma 1, we must have either (1) $O_i^t \not\perp\!\!\!\perp O_j^t|(\boldsymbol{W}^t \setminus \boldsymbol{Z}^t,\boldsymbol{S}^t)$ at some $t \in S(f(T))$ with $\boldsymbol{\lambda}_{\mathcal{I}_{O_i|\boldsymbol{W},\boldsymbol{S}}} \perp\!\!\!\perp \boldsymbol{\lambda}_{\mathcal{I}_{O_j|\boldsymbol{W},\boldsymbol{S}}}$, or (2) $O_i^t \not\perp\!\!\!\perp O_j^t|(\boldsymbol{W}^t \setminus \boldsymbol{Z}^t,\boldsymbol{S}^t)$ at some $t \in S(f(T))$ with $\boldsymbol{\lambda}_{\mathcal{I}_{O_i|\boldsymbol{W},\boldsymbol{S}}} \not\perp\!\!\!\perp \boldsymbol{\lambda}_{\mathcal{I}_{O_j|\boldsymbol{W},\boldsymbol{S}}}$. In either case, $O_i^t \not\perp\!\!\!\perp O_j^t|(\boldsymbol{W}^t \setminus \boldsymbol{Z}^t,\boldsymbol{S}^t)$ at some $t \in S(f(T))$, so the conclusion follows by the global directed Markov property. □

I find the above proposition useful for justifying part of the first orientation rule:

Lemma 8. *Assume mixture faithfulness, d-separation faithfulness with respect to* $\mathbb{P}_{\boldsymbol{X}^t}, \forall t \in S(f(T))$ *as well as parameter faithfulness. If* $O_i \perp\!\!\!\perp O_k|(\boldsymbol{W},\boldsymbol{S})$ *with minimal independence set* $\boldsymbol{W}$ *and* $O_j \in \boldsymbol{W}$*, then* $O_j^t \in \boldsymbol{An}(\{O_i^t, O_k^t\} \cup \boldsymbol{S}^t)$ *at some* $t \in S(f(T))$*. If further* $f(O_i, O_k|\boldsymbol{B},\boldsymbol{S})$ *is stationary for all* $\boldsymbol{B} \subseteq (\boldsymbol{W} \setminus O_j)$*, then* $O_j^t \in \boldsymbol{An}(\{O_i^t, O_k^t\} \cup \boldsymbol{S}^t)$ *at all* $t \in S(f(T))$*.*

Proof. I first introduce the following technical lemma which is a modification of Lemma 14 in [Spirtes et al., 1999]:

Lemma 9. *Suppose that we have $O_j^t \notin \boldsymbol{An}(\{O_i^t, O_k^t\} \cup \boldsymbol{S}^t)$ at all $t \in S(f(T))$. If there is a set $\boldsymbol{W} \subseteq \boldsymbol{O} \setminus \{O_i, O_k\}$ containing O_j such that $O_i^{t'}$ and $O_k^{t'}$ are d-connected given $\boldsymbol{B}^{t'} \cup \boldsymbol{S}^{t'}$ at some $t' \in S(f(T))$ for each subset $\boldsymbol{B} \subseteq (\boldsymbol{W} \setminus O_j)$, then $O_i^{t'}$ and $O_k^{t'}$ are also d-connected given $\boldsymbol{W}^{t'} \cup \boldsymbol{S}^{t'}$ for each subset $\boldsymbol{B} \subseteq (\boldsymbol{W} \setminus O_j)$; note that t' may not necessarily be the same for all $\boldsymbol{B} \subseteq (\boldsymbol{W} \setminus O_j)$.*

Proof. I write $G \in \boldsymbol{An}(\{O_i, O_k\} \cup \boldsymbol{S})$ if and only if $G^t \in \boldsymbol{An}(\{O_i^t, O_k^t\} \cup \boldsymbol{S}^t)$ at some time $t \in S(f(T))$. Let $\boldsymbol{B}_* = \boldsymbol{An}(\{O_i, O_k\} \cup \boldsymbol{S}) \cap \boldsymbol{W}$ and $\boldsymbol{B}_*^t = \boldsymbol{An}(\{O_i^t, O_k^t\} \cup \boldsymbol{S}^t) \cap \boldsymbol{W}^t$. Recall that $O_j \notin \boldsymbol{An}(\{O_i, O_k\} \cup \boldsymbol{S})$ by hypothesis, so $\boldsymbol{B}_* \subseteq (\boldsymbol{W} \setminus O_j)$. We also know that there exists a path $\pi^{t'}$ which d-connects $O_i^{t'}$ and $O_k^{t'}$ given $\boldsymbol{B}_*^{t'} \cup \boldsymbol{S}^{t'}$ by hypothesis. Thus, every vertex on $\pi^{t'}$ is in $\boldsymbol{An}(\{O_i^{t'}, O_k^{t'}\} \cup \boldsymbol{B}_*^{t'} \cup \boldsymbol{S}^{t'})$ by the definition of d-connection; in other words, every vertex on $\pi^{t'}$ has a corresponding vertex in $\boldsymbol{An}(\{O_i, O_k\} \cup \boldsymbol{B}_* \cup \boldsymbol{S})$. But since we also have $\boldsymbol{B}_*^{t'} = \boldsymbol{An}(\{O_i^{t'}, O_k^{t'}\} \cup \boldsymbol{S}^{t'}) \cap \boldsymbol{W}^{t'}$, every vertex on $\pi^{t'}$ more specifically has a vertex in $\boldsymbol{An}(\{O_i^{t'}, O_k^{t'}\} \cup \boldsymbol{B}_*^{t'} \cup \boldsymbol{S}^{t'}) = \boldsymbol{An}(\{O_i^{t'}, O_k^{t'}\} \cup \boldsymbol{S}^{t'})$ and therefore a corresponding vertex in $\boldsymbol{An}(\{O_i, O_k\} \cup \boldsymbol{S})$. Now observe that $\boldsymbol{W} \setminus \boldsymbol{B}_* = \boldsymbol{W} \cap (\neg \boldsymbol{An}(\{O_i, O_k\} \cup \boldsymbol{S}) \cup \neg \boldsymbol{W})) = \big((\boldsymbol{W} \cap \neg \boldsymbol{An}(\{O_i, O_k\} \cup \boldsymbol{S})) \cup (\boldsymbol{W} \cap \neg \boldsymbol{W}))\big) = \boldsymbol{W} \cap \neg \boldsymbol{An}(\{O_i, O_k\} \cup \boldsymbol{S})$, which is not in $\boldsymbol{An}(\{O_i, O_k\} \cup \boldsymbol{S})$. Thus, no vertex in $\boldsymbol{W} \setminus \boldsymbol{B}_*$ can exist on $\pi^{t'}$. Next, observe that $(\boldsymbol{B}_* \cup \boldsymbol{S}) \subseteq (\boldsymbol{W} \cup \boldsymbol{S})$ and $(\boldsymbol{W} \cup \boldsymbol{S}) \setminus (\boldsymbol{B}_* \cup \boldsymbol{S}) = \boldsymbol{W} \setminus \boldsymbol{B}_*$, so the additional vertices in $\boldsymbol{W} \cup \boldsymbol{S}$ cannot exist on $\pi^{t'}$. Hence $\pi^{t'}$ still d-connects $O_i^{t'}$ and $O_k^{t'}$ given $\boldsymbol{W}^{t'} \cup \boldsymbol{S}^{t'}$.

□

Now suppose for a contradiction that we have $O_j^t \notin \boldsymbol{An}(\{O_i^t, O_k^t\} \cup \boldsymbol{S}^t)$ at all $t \in S(f(T))$ for the first claim. With parameter faithfulness, we know that we have $O_i^t \not\perp\!\!\!\perp_d O_k^t | (\boldsymbol{B}^t, \boldsymbol{S}^t)$ at some $t \in S(f(T))$ for each $\boldsymbol{B} \subseteq (\boldsymbol{W} \setminus O_j)$ by Proposition 5. We can therefore invoke Lemma 9 and claim that we have $O_i^t \not\perp\!\!\!\perp_d O_k^t | (\boldsymbol{W}^t \cup \boldsymbol{S}^t)$ at some $t \in S(f(T))$. Hence, $O_i^t \not\perp\!\!\!\perp O_k^t | (\boldsymbol{W}^t, \boldsymbol{S}^t)$ at some $t \in S(f(T))$ by d-separation faithfulness and $O_i \not\perp\!\!\!\perp O_k | (\boldsymbol{W}, \boldsymbol{S})$ by the contrapositive of Lemma 4. However, this contradicts the fact that we must have $O_i \perp\!\!\!\perp O_k | (\boldsymbol{W}, \boldsymbol{S})$.

We will need the following lemma for the second claim:

Lemma 10. *Suppose that we have $O_j^{t'} \notin \boldsymbol{An}(\{O_i^{t'}, O_k^{t'}\} \cup \boldsymbol{S}^{t'})$ at some $t' \in S(f(T))$. If there is a set $\boldsymbol{W} \subseteq \boldsymbol{O} \setminus \{O_i, O_k\}$ containing O_j such that, for every subset $\boldsymbol{B} \subseteq (\boldsymbol{W} \setminus O_j)$, O_i^t and*

O_k^t are d-connected given $\boldsymbol{B}^t \cup \boldsymbol{S}^t$ at all $t \in S(f(T))$, then $O_i^{t'}$ and $O_k^{t'}$ are also d-connected given $\boldsymbol{W}^{t'} \cup \boldsymbol{S}^{t'}$.

Proof. The proof is similar to Lemma 9. Recall that $O_j^{t'} \notin \boldsymbol{An}(\{O_i^{t'}, O_k^{t'}\} \cup \boldsymbol{S}^{t'})$ by hypothesis, so $\boldsymbol{B}_*^{t'} \subseteq \boldsymbol{W}^{t'} \setminus O_j^{t'}$. We know that there exists a path $\pi^{t'}$ that d-connects $O_i^{t'}$ and $O_k^{t'}$ given $\boldsymbol{B}_*^{t'} \cup \boldsymbol{S}^{t'}$ by hypothesis. Thus, every vertex on $\pi^{t'}$ is in $\boldsymbol{An}(\{O_i^{t'}, O_k^{t'}\} \cup \boldsymbol{B}_*^{t'} \cup \boldsymbol{S}^{t'})$ by the definition of d-connection. But since we also have $\boldsymbol{B}_*^{t'} = \boldsymbol{An}(\{O_i^{t'}, O_k^{t'}\} \cup \boldsymbol{S}^{t'}) \cap \boldsymbol{W}^{t'}$, every vertex on $\pi^{t'}$ more specifically has a vertex in $\boldsymbol{An}(\{O_i^{t'}, O_k^{t'}\} \cup \boldsymbol{B}_*^{t'} \cup \boldsymbol{S}^{t'}) = \boldsymbol{An}(\{O_i^{t'}, O_k^{t'}\} \cup \boldsymbol{S}^{t'})$ and therefore a corresponding vertex in $\boldsymbol{An}(\{O_i, O_k\} \cup \boldsymbol{S})$. Now observe again that $\boldsymbol{W} \setminus \boldsymbol{B}_* = \boldsymbol{W} \cap \neg\boldsymbol{An}(\{O_i, O_k\} \cup \boldsymbol{S})$, which is not in $\boldsymbol{An}(\{O_i, O_k\} \cup \boldsymbol{S})$. Thus, no vertex in $\boldsymbol{W} \setminus \boldsymbol{B}_*$ can exist on $\pi^{t'}$. Again, $(\boldsymbol{W} \cup \boldsymbol{S}) \setminus (\boldsymbol{B}_* \cup \boldsymbol{S}) = \boldsymbol{W} \setminus \boldsymbol{B}_*$, so the additional vertices in $\boldsymbol{W} \cup \boldsymbol{S}$ cannot exist on $\pi^{t'}$. Hence $\pi^{t'}$ still d-connects $O_i^{t'}$ and $O_k^{t'}$ given $\boldsymbol{W}^{t'} \cup \boldsymbol{S}^{t'}$.

□

Now suppose again for a contradiction that $f(O_i, O_k|\boldsymbol{B}, \boldsymbol{S})$ is stationary for all $\boldsymbol{B} \subseteq (\boldsymbol{W} \setminus O_j)$, but we have $O_j^t \notin \boldsymbol{An}(\{O_i^t, O_k^t\} \cup \boldsymbol{S}^t)$ at some $t \in S(f(T))$. We can then follow a similar deductive argument as in the proof of the previous claim. With parameter faithfulness, we know that we have $O_i^t \not\perp\!\!\!\perp_d O_k^t|(\boldsymbol{B}^t \cup \boldsymbol{S}^t)$ at some $t \in S(f(T))$ for each $\boldsymbol{B} \subseteq (\boldsymbol{W} \setminus O_j)$ by Proposition 5. It follows that $O_i^t \not\perp\!\!\!\perp O_k^t|(\boldsymbol{B}^t, \boldsymbol{S}^t)$ at some $t \in S(f(T))$ for each $\boldsymbol{B} \subseteq (\boldsymbol{W} \setminus O_j)$ by d-separation faithfulness. We can more strongly claim that we have $O_i^t \not\perp\!\!\!\perp O_k^t|(\boldsymbol{B}^t, \boldsymbol{S}^t)$ at all $t \in S(f(T))$ for each $\boldsymbol{B} \subseteq (\boldsymbol{W} \setminus O_j)$ by Lemma 2, because we also know that $f(O_i, O_k|\boldsymbol{B}, \boldsymbol{S})$ is stationary for all $\boldsymbol{B} \subseteq (\boldsymbol{W} \setminus O_j)$. Hence, $O_i^t \not\perp\!\!\!\perp_d O_k^t|(\boldsymbol{B}^t, \boldsymbol{S}^t)$ at all $t \in S(f(T))$ for each $\boldsymbol{B} \subseteq (\boldsymbol{W} \setminus O_j)$ by the global directed Markov property. We can therefore invoke Lemma 10 and claim that we have $O_i^t \not\perp\!\!\!\perp_d O_k^t|(\boldsymbol{W}^t, \boldsymbol{S}^t)$ at some $t \in S(f(T))$. Thus, $O_i^t \not\perp\!\!\!\perp O_k^t|(\boldsymbol{W}^t, \boldsymbol{S}^t)$ at some $t \in S(f(T))$ by d-separation faithfulness and $O_i \not\perp\!\!\!\perp O_k|(\boldsymbol{W}, \boldsymbol{S})$ by the contrapositive of Lemma 4. However, this again contradicts the fact that we must have $O_i \perp\!\!\!\perp O_k|(\boldsymbol{W}, \boldsymbol{S})$.

□

I can therefore use the above lemma in order to apply the following orientation rules:

1. R1a$_*$: If (1) $O_i * \!\!\rightarrow O_j \circ\!\!-\!\!* O_k$ or $O_i * \!\!\rightarrow O_j \bullet\!\!-\!\!* O_k$, (2) $O_i \perp\!\!\!\perp O_k|(\boldsymbol{W}, \boldsymbol{S})$ with minimal

independence set $\boldsymbol{W}$ and $O_j \in \boldsymbol{W}$, and (3) $f(O_i, O_k|\boldsymbol{B}, \boldsymbol{S})$ is stationary for all $\boldsymbol{B} \subseteq (\boldsymbol{W} \setminus O_j)$, then orient $O_i{*\!\!\rightarrow}\, O_j \circ\!\!-\!\!* O_k$ or $O_i{*\!\!\rightarrow}\, O_j \bullet\!\!-\!\!* O_k$ as $O_i{*\!\!\rightarrow}\, O_j —\!\!* O_k$

2. R1b$_*$: If (1) $O_i{*\!\!\rightarrow}\, O_j \circ\!\!-\!\!* O_k$ or $O_i{*\!\!\rightarrow}\, O_j \blacktriangleleft\!\!-\!\!* O_k$, (2) $O_i \perp\!\!\!\perp O_k|(\boldsymbol{W}, \boldsymbol{S})$ with minimal independence set $\boldsymbol{W}$ and $O_j \in \boldsymbol{W}$, and (3) $f(O_i, O_k|\boldsymbol{B}, \boldsymbol{S})$ is non-stationary for at least one $\boldsymbol{B} \subseteq (\boldsymbol{W} \setminus O_j)$, then orient $O_i{*\!\!\rightarrow}\, O_j \circ\!\!-\!\!* O_k$ or $O_i{*\!\!\rightarrow}\, O_j \blacktriangleleft\!\!-\!\!* O_k$ as $O_i{*\!\!\rightarrow}\, O_j \bullet\!\!-\!\!* O_k$.

3. R1c$_*$: If (1) $O_i *\!\!\rightarrow O_j \circ\!\!-\!\!* O_k$, $O_i *\!\!\rightarrow O_j \bullet\!\!-\!\!* O_k$, $O_i *\!\!-\!\!\blacksquare O_j \circ\!\!-\!\!* O_k$ or $O_i *\!\!-\!\!\blacksquare O_j \bullet\!\!-\!\!* O_k$, (2) $O_i \perp\!\!\!\perp O_k|(\boldsymbol{W}, \boldsymbol{S})$ with minimal independence set $\boldsymbol{W}$ and $O_j \in \boldsymbol{W}$, and (3) $f(O_i, O_k|\boldsymbol{B}, \boldsymbol{S})$ is stationary for all $\boldsymbol{B} \subseteq \boldsymbol{W} \setminus O_j$, then orient (C1) $O_i *\!\!\rightarrow O_j —\!\!* O_k$ when we have $O_i *\!\!\rightarrow O_j \circ\!\!-\!\!* O_k$ or $O_i *\!\!\rightarrow O_j \bullet\!\!-\!\!* O_k$, or (C2) $O_i *\!\!-\!\!\blacksquare O_j —\!\!* O_k$ when we have $O_i *\!\!-\!\!\blacksquare O_j \circ\!\!-\!\!* O_k$ or $O_i *\!\!-\!\!\blacksquare O_j \bullet\!\!-\!\!* O_k$.

We also have the following result:

Lemma 11. *If we have the edges $O_i —\!\!\blacksquare O_j$ or $O_i \blacksquare\!\!— O_j$, then we cannot have an incoming unfilled arrowhead at O_i or O_j. Similarly, if we have the edge $O_i — O_j$, then we cannot have an incoming unfilled arrowhead, filled arrowhead or square at O_i or O_j.*

Proof. Recall that we do not allow instantaneous feedback loops. Thus, if we have $O_i —\!\!\blacksquare O_j$ or $O_i \blacksquare\!\!— O_j$, then both O_i^t and O_j^t must be ancestors of $\boldsymbol{S}^t$ at some time point $t \in S(f(T))$. Hence, O_i^t and O_j^t cannot be non-ancestors of $\boldsymbol{S}^t$ at all time points $t \in S(f(T))$. Similarly, if we have $O_i — O_j$, then both O_i^t and O_j^t must be ancestors of $\boldsymbol{S}^t$ at all time points $t \in S(f(T))$. Hence, O_i^t and O_j^t cannot be non-ancestors of $\boldsymbol{S}^t$ at some time point $t \in S(f(T))$. □

We can thus further expand on R1 as follows by taking the contrapositive of Lemma 11:

1. R1a: If (1)-(3) hold as in R1a$_*$, then orient $O_i{*\!\!\rightarrow}\, O_j \circ\!\!-\!\!* O_k$ or $O_i{*\!\!\rightarrow}\, O_j \bullet\!\!-\!\!* O_k$ as $O_i{*\!\!\rightarrow}\, O_j \rightarrow O_k$.

2. R1b: If (1)-(3) hold as in R1b$_*$, then orient $O_i{*\!\!\rightarrow}\, O_j \circ\!\!-\!\!* O_k$ or $O_i{*\!\!\rightarrow}\, O_j \blacktriangleleft\!\!-\!\!* O_k$ as $O_i{*\!\!\rightarrow}\, O_j \bullet\!\!\rightarrow O_k$.

3. R1c: If (1)-(3) hold as in R1c$_*$, then orient (C1) $O_i *\!\!\rightarrow O_j \rightarrow O_k$ when $O_i *\!\!\rightarrow O_j \circ\!\!-\!\!* O_k$ or $O_i *\!\!\rightarrow O_j \bullet\!\!-\!\!* O_k$, or (C2) $O_i *\!\!-\!\!\blacksquare O_j \rightarrow O_k$ when $O_i *\!\!-\!\!\blacksquare O_j \circ\!\!-\!\!* O_k$ or $O_i *\!\!-\!\!\blacksquare O_j \bullet\!\!-\!\!* O_k$.

5.2.6 Fifth, Ninth and Tenth Orientation Rules

Let us now tackle three other rules: R5, R9 and R10. We will see that R5 is actually not required. The proofs of the rules follow by establishing contradictions with R1 by the concept of an uncovered path:

Definition 8. *An uncovered path $\pi = \langle O_0, \ldots, O_n \rangle$ is a path where O_{i-1} and O_{i+1} are non-adjacent for every $1 \leq i \leq n-1$. In other words, every consecutive triple on π is unshielded.*

I say that an uncovered path is stationary when, for every $O_{i-1} \perp\!\!\!\perp O_{i+1} | (\boldsymbol{W}, \boldsymbol{S})$ with minimal independence set $\boldsymbol{W}$, $f(O_{i-1}, O_{i+1} | \boldsymbol{B}, \boldsymbol{S})$ is stationary for all $\boldsymbol{B} \subseteq (\boldsymbol{W} \setminus O_i)$. Likewise, I say that an uncovered path is non-stationary when the former description holds but there exists one and only one conditional independence relation $O_{i-1} \perp\!\!\!\perp O_{i+1} | (\boldsymbol{W}, \boldsymbol{S})$ with minimal independence set $\boldsymbol{W}$ such that $f(O_{i-1}, O_{i+1} | \boldsymbol{B}, \boldsymbol{S})$ is non-stationary for some $\boldsymbol{B} \subseteq (\boldsymbol{W} \setminus O_i)$.

Here are rules 9 and 10:

Lemma 12. *Assume mixture faithfulness, d-separation faithfulness with respect to $\mathbb{P}_{\boldsymbol{X}^t}, \forall t \in S(f(T))$ as well as parameter faithfulness. The following variations of FCI's R9 are sound:*

1. *R9a: If (1) $O_i \circ\!\!-\!\!* O_k$ or $O_i \bullet\!\!-\!\!* O_k$, (2) $\pi = \langle O_i, O_j, O_l, \ldots, O_k \rangle$ is a stationary uncovered path, (3) $O_k \perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S})$ with minimal independence set $\boldsymbol{W}$, $O_i \in \boldsymbol{W}$, and (4) $f(O_k, O_j | \boldsymbol{B}, \boldsymbol{S})$ is stationary for all $\boldsymbol{B} \subseteq (\boldsymbol{W} \setminus O_i)$, then orient $O_i \circ\!\!-\!\!* O_k$ or $O_i \bullet\!\!-\!\!* O_k$ as $O_i \!-\!\!* O_k$.*
2. *R9b: If (1) $O_i \circ\!\!-\!\!* O_k$ or $O_i \leftarrow\!\!* O_k$, (2) $\pi = \langle O_i, O_j, O_l, \ldots, O_k \rangle$ is a non-stationary uncovered path, (3) $O_k \perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S})$ with minimal independence set $\boldsymbol{W}$, $O_i \in \boldsymbol{W}$, and (4) $f(O_k, O_j | \boldsymbol{B}, \boldsymbol{S})$ is stationary for all $\boldsymbol{B} \subseteq (\boldsymbol{W} \setminus O_i)$, then orient $O_i \circ\!\!-\!\!* O_k$ or $O_i \leftarrow\!\!* O_k$ as $O_i \bullet\!\!-\!\!* O_k$.*
3. *R9c: If (1) $O_i \circ\!\!-\!\!* O_k$ or $O_i \leftarrow\!\!* O_k$, (2) $\pi = \langle O_i, O_j, O_l, \ldots, O_k \rangle$ is a stationary uncovered path, (3) $O_k \perp\!\!\!\perp O_j | (\boldsymbol{W}, \boldsymbol{S})$ with minimal independence set $\boldsymbol{W}$, $O_i \in \boldsymbol{W}$, and (4) $f(O_k, O_j | \boldsymbol{B}, \boldsymbol{S})$ is non-stationary for some $\boldsymbol{B} \subseteq (\boldsymbol{W} \setminus O_i)$, then orient $O_i \circ\!\!-\!\!* O_k$ or $O_i \leftarrow\!\!* O_k$ as $O_i \bullet\!\!-\!\!* O_k$.*

Proof. R9a: Suppose that we instead have an unfilled arrowhead or a square at O_i for $O_i \circ\!\!-\!\!* O_k$

or $O_i \bullet\!\!-\!\!* O_k$ (we can only have a square for the latter). Then we can iteratively apply R1c on π until the transitivity of the added unfilled tails contradicts the unfilled arrowhead or square at O_i.

R9b: Suppose that we instead have an unfilled arrowhead at O_i for $O_i \circ\!\!-\!\!* O_k$ or $O_i \blacktriangleleft\!\!-\!\!* O_k$. Then we can apply R1a iteratively until we encounter a non-stationary density. We can then apply R1b once at the non-stationary density, and finally R1c iteratively on π until the transitivity of the added tails contradicts the unfilled arrowhead at O_i.

R9c: Suppose that we instead have an unfilled arrowhead at O_i for $O_i \circ\!\!-\!\!* O_k$ or $O_i \blacktriangleleft\!\!-\!\!* O_k$. Then we can apply R1b once at the non-stationary density and then R1c iteratively on π until the transitivity of the added tails contradicts the unfilled arrowhead at O_i. □

Note that FCI's R5 is just a specific instance of R9a. Thus, we actually will *not* include variants of R5 in the proposed algorithm, since they are covered by the proposed R9. Recall however that FCI's R5 is required in the original FCI algorithm, because FCI's R9 is less general than the proposed R9.[1] We will keep the same numbering of orientation rules despite dropping R5, so we can easily map back and forth between the proposed algorithm's rules and those of FCI.

We now have R10 whose argument proceeds similar to that of R9:

Lemma 13. *Assume mixture faithfulness, d-separation faithfulness with respect to* $\mathbb{P}_{\boldsymbol{X}^t}, \forall t \in S(f(T))$ *as well as parameter faithfulness. The following variations of FCI's R10 are sound:*

1. *R10a: Suppose (1)* $O_i \circ\!\!-\!\!* O_k$ *or* $O_i \bullet\!\!-\!\!* O_k$*, (2)* $O_j \!-\!\!* O_k *\!\!- O_l$*, and (3)* π_1 *is a stationary uncovered path from* O_i *to* O_j*, and* π_2 *is a stationary uncovered path from* O_i *to* O_l*. Let* O_m *be a vertex adjacent to* O_i *on* π_1 *(*O_m *could be* O_j*) and* O_n *be the vertex adjacent to* O_i *on* π_2 *(*O_n *could be* O_l*). If (4)* O_m *and* O_n *are distinct, (5)* $O_m \perp\!\!\!\perp O_n | (\boldsymbol{W}, \boldsymbol{S})$ *with minimal independence set* $\boldsymbol{W}$*,* $O_i \in \boldsymbol{W}$ *and (6)* $f(O_m, O_n | \boldsymbol{B}, \boldsymbol{S})$ *is stationary for all* $\boldsymbol{B} \subseteq (\boldsymbol{W} \setminus O_i)$*, then orient* $O_i \circ\!\!-\!\!* O_k$ *or* $O_i \bullet\!\!-\!\!* O_k$ *as* $O_i \!-\!\!* O_k$*.*
2. *R10b: Suppose (1)* $O_i \circ\!\!-\!\!* O_k$ *or* $O_i \blacktriangleleft\!\!-\!\!* O_k$*, (2)* $O_j \bullet\!\!-\!\!* O_k *\!\!- O_l$*,* $O_j \!-\!\!* O_k *\!\!-\!\!\bullet O_l$*,* $O_j \blacksquare\!\!-\!\!* O_k *\!\!- O_l$ *or* $O_j \!-\!\!* O_k *\!\!-\!\!\blacksquare O_l$ *and (3)* π_1 *is an stationary uncovered path from* O_i *to*

[1] FCI's R9 requires an arrowhead at O_k and an uncovered potentially directed path, whereas the proposed R9 only requires an asterisk and an uncovered path.

O_j, and π_2 is an stationary uncovered path from O_i to O_l. Let O_m be a vertex adjacent to O_i on π_1 (O_m could be O_j) and O_n be the vertex adjacent to O_i on π_2 (O_n could be O_l). If (4) O_m and O_n are distinct, (5) $O_m \perp\!\!\!\perp O_n|(\boldsymbol{W}, \boldsymbol{S})$ with minimal independence set $\boldsymbol{W}$, $O_i \in \boldsymbol{W}$ and (6) $f(O_m, O_n|\boldsymbol{B}, \boldsymbol{S})$ is stationary for all $\boldsymbol{B} \subseteq (\boldsymbol{W} \setminus O_i)$, then orient $O_i \circ\!\!-\!\!* O_k$ or $O_i \leftarrow\!\!* O_k$ as $O_i \bullet\!\!-\!\!* O_k$.

3. *R10c: Suppose (1) $O_i \circ\!\!-\!\!* O_k$ or $O_i \leftarrow\!\!* O_k$, (2) $O_j \bullet\!\!-\!\!* O_k *\!\!-\!\!\bullet O_l$, $O_j \blacksquare\!\!-\!\!* O_k *\!\!-\!\!\bullet O_l$, $O_j \bullet\!\!-\!\!* O_k *\!\!-\!\!\blacksquare O_l$ or $O_j \blacksquare\!\!-\!\!* O_k *\!\!-\!\!\blacksquare O_l$ are contemporaneous with respect to $O_j -\!\!* O_k *\!\!- O_l$, and (3) π_1 is a stationary uncovered path from O_i to O_j, and π_2 is a stationary uncovered path from O_i to O_l. Let O_m be a vertex adjacent to O_i on π_1 (O_m could be O_j) and O_n be the vertex adjacent to O_i on π_2 (O_n could be O_l). If (4) O_m and O_n are distinct, (5) $O_m \perp\!\!\!\perp O_n|(\boldsymbol{W}, \boldsymbol{S})$ with minimal independence set $\boldsymbol{W}$, $O_i \in \boldsymbol{W}$ and (6) $f(O_m, O_n|\boldsymbol{B}, \boldsymbol{S})$ is stationary for all $\boldsymbol{B} \subseteq (\boldsymbol{W} \setminus O_i)$, then orient $O_i \circ\!\!-\!\!* O_k$ or $O_i \leftarrow\!\!* O_k$ as $O_i \bullet\!\!-\!\!* O_k$.*
4. *R10d: Suppose (1) $O_i \circ\!\!-\!\!* O_k$ or $O_i \leftarrow\!\!* O_k$, (2) $O_j -\!\!* O_k *\!\!- O_l$, and (3) π_1 is an uncovered path from O_i to O_j, and π_2 is an uncovered path from O_i to O_l with either π_1 or π_2 non-stationary. Let O_m be a vertex adjacent to O_i on π_1 (O_m could be O_j) and O_n be the vertex adjacent to O_i on π_2 (O_n could be O_l). If (4) O_m and O_n are distinct, (5) $O_m \perp\!\!\!\perp O_n|(\boldsymbol{W}, \boldsymbol{S})$ with minimal independence set $\boldsymbol{W}$, $O_i \in \boldsymbol{W}$ and (6) $f(O_m, O_n|\boldsymbol{B}, \boldsymbol{S})$ is stationary for all $\boldsymbol{B} \subseteq (\boldsymbol{W} \setminus O_i)$, then orient $O_i \circ\!\!-\!\!* O_k$ or $O_i \leftarrow\!\!* O_k$ as $O_i \bullet\!\!-\!\!* O_k$.*
5. *R10e: Suppose (1) $O_i \circ\!\!-\!\!* O_k$ or $O_i \leftarrow\!\!* O_k$, (2) $O_j -\!\!* O_k *\!\!- O_l$, (3) π_1 is a stationary uncovered path from O_i to O_j, and π_2 is a stationary uncovered path from O_i to O_l. Let O_m be a vertex adjacent to O_i on π_1 (O_m could be O_j) and O_n be the vertex adjacent to O_i on π_2 (O_n could be O_l). If (4) O_m and O_n are distinct, (5) $O_m \perp\!\!\!\perp O_n|(\boldsymbol{W}, \boldsymbol{S})$ with minimal independence set $\boldsymbol{W}$, $O_i \in \boldsymbol{W}$ and (6) $f(O_m, O_n|\boldsymbol{B}, \boldsymbol{S})$ is non-stationary for some $\boldsymbol{B} \subseteq (\boldsymbol{W} \setminus O_i)$, then orient $O_i \circ\!\!-\!\!* O_k$ or $O_i \leftarrow\!\!* O_k$ as $O_i \bullet\!\!-\!\!* O_k$.*

Proof. R10a: Assume an unfilled arrowhead or a square at O_i. Then, we can successively apply R1c along both π_1 and π_2. In either case, O_i is an ancestor of $O_k \cup \boldsymbol{S}$ at all $t \in S(f(T))$ by transitivity of the added unfilled tails which contradicts the unfilled arrowhead or square at O_i.

R10b,c: Assume an unfilled arrowhead at O_i. Then we can successively apply R1a along

both π_1 and π_2. Here, we arrive at the contradiction that O_i is an ancestor of $O_k \cup \boldsymbol{S}$ at some $t \in S(f(T))$ by transitivity of the added tails.

R10d,e: Assume an unfilled arrowhead at O_i. Then we can successively apply R1a along both π_1 and π_2 or R1b when we encounter the non-stationary density. We then arrive at the contradiction that O_i is an ancestor of $O_k \cup \boldsymbol{S}$ at some $t \in S(f(T))$ by transitivity of the added tails.

□

5.2.7 Remaining Orientation Rules

Let us now wrap-up the easier rules. I enumerate the sound variations of FCI's rules R2-R3 and R6-R8.

Lemma 14. *The following variations of FCI's R2 are sound:*

1. *R2a: If (1) $O_i \ast\!\!\rightarrow O_j \rightarrow\!\!\ast O_k$ or $O_i \rightarrow\!\!\ast O_j \ast\!\!\rightarrow O_k$, and (2) $O_i \ast\!\!-\!\!\circ O_k$ or $O_i \ast\!\!\rightarrow O_k$, then orient $O_i \ast\!\!-\!\!\circ O_k$ or $O_i \ast\!\!\rightarrow O_k$ as $O_i \ast\!\!\rightarrow O_k$.*
2. *R2b: If (1) $O_i \ast\!\!\rightarrow O_j \rightarrow\!\!\ast O_k$, $O_i \ast\!\!-\!\!\blacksquare O_j \rightarrow\!\!\ast O_k$, $O_i \rightarrow\!\!\ast O_j \ast\!\!\rightarrow O_k$ or $O_i \rightarrow\!\!\ast O_j \ast\!\!-\!\!\blacksquare O_k$ and (2) $O_i \ast\!\!-\!\!\circ O_k$ or $O_i \ast\!\!-\!\!\bullet O_k$, then orient $O_i \ast\!\!-\!\!\circ O_k$ or $O_i \ast\!\!-\!\!\bullet O_k$ as $O_i \ast\!\!\rightarrow O_k$.*
3. *R2c: If (1) $O_i \ast\!\!\rightarrow O_j \bullet\!\!-\!\!\ast O_k$, $O_i \ast\!\!\rightarrow O_j \blacksquare\!\!-\!\!\ast O_k$, $O_i \bullet\!\!-\!\!\ast O_j \ast\!\!\rightarrow O_k$ or $O_i \blacksquare\!\!-\!\!\ast O_j \ast\!\!\rightarrow O_k$ and (2) $O_i \ast\!\!-\!\!\circ O_k$ or $O_i \ast\!\!-\!\!\bullet O_k$, then orient $O_i \ast\!\!-\!\!\circ O_k$ or $O_i \ast\!\!-\!\!\bullet O_k$ as $O_i \ast\!\!\rightarrow O_k$.*
4. *R2d: If (1) $O_i \ast\!\!\rightarrow O_j \bullet\!\!-\!\!\ast O_k$, $O_i \ast\!\!-\!\!\blacksquare O_j \bullet\!\!-\!\!\ast O_k$, $O_i \ast\!\!\rightarrow O_j \blacksquare\!\!-\!\!\ast O_k$ or $O_i \ast\!\!-\!\!\blacksquare O_j \blacksquare\!\!-\!\!\ast O_k$ is contemporaneous with respect to $O_i \ast\!\!\rightarrow O_j \rightarrow\!\!\ast O_k$ and (2) $O_i \ast\!\!-\!\!\circ O_k$ or $O_i \ast\!\!-\!\!\bullet O_k$, then orient $O_i \ast\!\!-\!\!\circ O_k$ or $O_i \ast\!\!-\!\!\bullet O_k$ as $O_i \ast\!\!\rightarrow O_k$.*
5. *R2e: If (1) $O_i \bullet\!\!-\!\!\ast O_j \ast\!\!\rightarrow O_k$, $O_i \blacksquare\!\!-\!\!\ast O_j \ast\!\!\rightarrow O_k$, $O_i \bullet\!\!-\!\!\ast O_j \ast\!\!-\!\!\blacksquare O_k$ or $O_i \blacksquare\!\!-\!\!\ast O_j \ast\!\!-\!\!\blacksquare O_k$ is contemporaneous with respect to $O_i \rightarrow\!\!\ast O_j \ast\!\!\rightarrow O_k$ and (2) $O_i \ast\!\!-\!\!\circ O_k$ or $O_i \ast\!\!-\!\!\bullet O_k$, then orient $O_i \ast\!\!-\!\!\circ O_k$ or $O_i \ast\!\!-\!\!\bullet O_k$ as $O_i \ast\!\!\rightarrow O_k$.*

Proof. R2a: Suppose to the contrary that we have an unfilled tail or square at O_k. If $O_i \ast\!\!\rightarrow O_j \rightarrow\!\!\ast O_k$, then $O_j^t \in \boldsymbol{An}(O_i^t \cup \boldsymbol{S}^t)$ for all or some $t \in S(f(T))$, respectively, which contradicts the arrowhead at O_j. If $O_i \rightarrow\!\!\ast O_j \ast\!\!\rightarrow O_k$, then $O_k^t \in \boldsymbol{An}(O_j^t \cup \boldsymbol{S}^t)$ for all or some $t \in S(f(T))$, respectively, which contradicts the arrowhead at O_k.

R2b: Suppose to the contrary that we have an unfilled tail at O_k. Then either $O_j^t \in \boldsymbol{An}(O_i^t \cup \boldsymbol{S}^t)$ or $O_k^t \in \boldsymbol{An}(O_j^t \cup \boldsymbol{S}^t)$ at all $t \in S(f(T))$ which contradicts the filled arrowhead at O_j or O_k, respectively.

R2c: Suppose to the contrary that we have an unfilled tail at O_k. Then either $O_j^t \in \boldsymbol{An}(O_i^t \cup \boldsymbol{S}^t)$ or $O_k^t \in \boldsymbol{An}(O_j^t \cup \boldsymbol{S}^t)$ at some $t \in S(f(T))$ which contradicts the unfilled arrowhead at O_j or O_k, respectively.

R2d: Suppose to the contrary that we have an unfilled tail at O_k. Then $O_j^t \in \boldsymbol{An}(O_i^t \cup \boldsymbol{S}^t)$ at some $t \in A \subseteq S(f(T))$ which contradicts the filled arrowhead at O_j for $t \in A$.

R2e: Suppose to the contrary that we have an unfilled tail at O_k. Then $O_k^t \in \boldsymbol{An}(O_j^t \cup \boldsymbol{S}^t)$ at some $t \in A \subseteq S(f(T))$ which contradicts the filled arrowhead at O_j for $t \in A$. □

We now move onto R3:

Lemma 15. *The following variations of FCI's R3 are sound under mixture faithfulness, d-separation faithfulness with respect to* $\mathbb{P}_{\boldsymbol{X}^t}, \forall t \in S(f(T))$ *as well as parameter faithfulness:*

1. *R3a: If (1)* $O_i * \!\!\rightarrow O_j \leftarrow\!\!* O_k$*, (2)* $O_i * \!\!-\!\!* O_l * \!\!-\!\!* O_k$*, (3)* $O_i \perp\!\!\!\perp O_k | (\boldsymbol{W}, \boldsymbol{S})$ *with* $O_l \in \boldsymbol{W}$ *and minimal independence set* $\boldsymbol{W}$*, (4)* $f(O_i, O_k | \boldsymbol{B}, \boldsymbol{S})$ *is stationary for all* $\boldsymbol{B} \subseteq (\boldsymbol{W} \setminus O_l)$*, and (5)* $O_l * \!\!-\!\!\circ O_j$ *or* $O_l * \!\!-\!\!\blacktriangleright O_j$*, then orient* $O_l * \!\!-\!\!\circ O_j$ *or* $O_l * \!\!-\!\!\blacktriangleright O_j$ *as* $O_l * \!\!\rightarrow O_j$*.*
2. *R3b: If (1)* $O_i * \!\!-\!\!\blacktriangleright O_j \leftarrow\!\!* O_k$*,* $O_i * \!\!\rightarrow O_j \blacktriangleleft\!\!-\!\!* O_k$*,* $O_i * \!\!-\!\!\blacksquare O_j \leftarrow\!\!* O_k$ *or* $O_i * \!\!\rightarrow O_j \blacksquare\!\!-\!\!* O_k$*, (2)* $O_i * \!\!-\!\!* O_l * \!\!-\!\!* O_k$*, (3)* $O_i \perp\!\!\!\perp O_k | (\boldsymbol{W}, \boldsymbol{S})$ *with* $O_l \in \boldsymbol{W}$ *and minimal independence set* $\boldsymbol{W}$*, (4)* $f(O_i, O_k | \boldsymbol{B}, \boldsymbol{S})$ *is stationary for all* $\boldsymbol{B} \subseteq (\boldsymbol{W} \setminus O_l)$*, and (5)* $O_l * \!\!-\!\!\circ O_j$ *or* $O_l * \!\!-\!\!\bullet O_j$*, then orient* $O_l * \!\!-\!\!\circ O_j$ *or* $O_l * \!\!-\!\!\bullet O_j$ *as* $O_l * \!\!-\!\!\blacktriangleright O_j$*.*
3. *R3c: If (1) we have the contemporaneous path* $O_i * \!\!-\!\!\blacktriangleright O_j \blacktriangleleft\!\!-\!\!* O_k$*,* $O_i * \!\!-\!\!\blacksquare O_j \blacktriangleleft\!\!-\!\!* O_k$*,* $O_i * \!\!-\!\!\blacktriangleright O_j \blacksquare\!\!-\!\!* O_k$ *or* $O_i * \!\!-\!\!\blacksquare O_j \blacksquare\!\!-\!\!* O_k$ *with respect to* $O_i * \!\!\rightarrow O_j \leftarrow\!\!* O_k$*, (2)* $O_i * \!\!-\!\!* O_l * \!\!-\!\!* O_k$*, (3)* $O_i \perp\!\!\!\perp O_k | (\boldsymbol{W}, \boldsymbol{S})$ *with* $O_l \in \boldsymbol{W}$ *and minimal independence set* $\boldsymbol{W}$*, (4)* $f(O_i, O_k | \boldsymbol{B}, \boldsymbol{S})$ *is stationary for all* $\boldsymbol{B} \subseteq (\boldsymbol{W} \setminus O_l)$*, and (5)* $O_l * \!\!-\!\!\circ O_j$ *or* $O_l * \!\!-\!\!\bullet O_j$*, then orient* $O_l * \!\!-\!\!\circ O_j$ *or* $O_l * \!\!-\!\!\bullet O_j$ *as* $O_l * \!\!-\!\!\blacktriangleright O_j$*.*
4. *R3d: If (1)* $O_i * \!\!\rightarrow O_j \leftarrow\!\!* O_k$*, (2)* $O_i * \!\!-\!\!* O_l * \!\!-\!\!* O_k$*, (3)* $O_i \perp\!\!\!\perp O_k | (\boldsymbol{W}, \boldsymbol{S})$ *with* $O_l \in \boldsymbol{W}$ *and minimal independence set* $\boldsymbol{W}$*, (4)* $f(O_i, O_k | \boldsymbol{B}, \boldsymbol{S})$ *is non-stationary for some* $\boldsymbol{B} \subseteq (\boldsymbol{W} \setminus O_l)$*, and (5)* $O_l * \!\!-\!\!\circ O_j$ *or* $O_l * \!\!-\!\!\bullet O_j$*, then orient* $O_l * \!\!-\!\!\circ O_j$ *or* $O_l * \!\!-\!\!\bullet O_j$ *as* $O_l * \!\!-\!\!\blacktriangleright O_j$*.*

Proof. R3a: Observe by (3) and (4) that $O_l^t \in \boldsymbol{An}(O_i^t \cup O_k^t \cup \boldsymbol{S}^t)$ at all $t \in S(f(T))$ due to Lemma 8. Assume to the contrary that we have an unfilled tail or square at O_j. But then $O_j^t \in \boldsymbol{An}(O_i^t \cup O_k^t \cup \boldsymbol{S}^t)$ at some $t \in S(f(T))$, which contradicts the unfilled arrowheads at O_j.

R3b: Observe by (3) and (4) that $O_l^t \in \boldsymbol{An}(O_i^t \cup O_k^t \cup \boldsymbol{S}^t)$ at all $t \in S(f(T))$ due to Lemma 8. Assume to the contrary that we have an unfilled tail at O_j. But then $O_j^t \in \boldsymbol{An}(O_i^t \cup O_k^t \cup \boldsymbol{S}^t)$ at all $t \in S(f(T))$, which contradicts the arrowheads at O_j.

R3c: Observe by (3) and (4) that $O_l^t \in \boldsymbol{An}(O_i^t \cup O_k^t \cup \boldsymbol{S}^t)$ at all $t \in S(f(T))$ due to Lemma 8. Assume to the contrary that we have an unfilled tail at O_j. But then $O_j^t \in \boldsymbol{An}(O_i^t \cup O_k^t \cup \boldsymbol{S}^t)$ at all $t \in S(f(T))$, which contradicts the contemporaneous filled arrowheads at O_j.

R3d: Observe by (3) and (4) that $O_l^t \in \boldsymbol{An}(O_i^t \cup O_k^t \cup \boldsymbol{S}^t)$ at some $t \in S(f(T))$ due to Lemma 8. Assume to the contrary that we have an unfilled tail at O_j. But then $O_j^t \in \boldsymbol{An}(O_i^t \cup O_k^t \cup \boldsymbol{S}^t)$ at some $t \in S(f(T))$, which contradicts the unfilled arrowheads at O_j.

□

Lemma 16. *The following variations of FCI's R6 are sound:*

1. *R6a: If $O_i - O_j \circ\!\!-\!\!* O_k$ or $O_i - O_j \bullet\!\!-\!\!* O_k$ (O_i and O_k may or may not be adjacent), then orient $O_j \circ\!\!-\!\!* O_k$ or $O_j \bullet\!\!-\!\!* O_k$ as $O_j \text{—}\!\!* O_k$.*
2. *R6b: If $O_i \text{—}\!\bullet O_j \circ\!\!-\!\!* O_k$, $O_i \text{—}\!\blacksquare O_j \circ\!\!-\!\!* O_k$, $O_i \text{—}\!\bullet O_j \leftarrow\!\!* O_k$ or $O_i \text{—}\!\blacksquare O_j \leftarrow\!\!* O_k$ (O_i and O_k may or may not be adjacent), then orient $O_j \circ\!\!-\!\!* O_k$ or $O_j \leftarrow\!\!* O_k$ as $O_j \bullet\!\!-\!\!* O_k$.*
3. *R6c: If $O_i \bullet\!\!-\!\!O_j \circ\!\!-\!\!* O_k$, $O_i \blacksquare\!\!- O_j \circ\!\!-\!\!* O_k$, $O_i \bullet\!\!-\!\!O_j \leftarrow\!\!* O_k$ or $O_i \blacksquare\!\!- O_j \leftarrow\!\!* O_k$ (O_i and O_k may or may not be adjacent), then orient $O_j \circ\!\!-\!\!* O_k$ or $O_j \leftarrow\!\!* O_k$ as $O_j \bullet\!\!-\!\!* O_k$.*
4. *R6d: If $O_i \bullet\!\!-\!\!\bullet O_j \circ\!\!-\!\!* O_k$, $O_i \blacksquare\!\!-\!\!\bullet O_j \circ\!\!-\!\!* O_k$, $O_i \bullet\!\!-\!\!\blacksquare O_j \circ\!\!-\!\!* O_k$ or $O_i \blacksquare\!\!-\!\!\blacksquare O_j \circ\!\!-\!\!* O_k$ is contemporaneous with respect to $O_i - O_j \circ\!\!-\!\!* O_k$ (O_i and O_k may or may not be adjacent), then orient $O_j \circ\!\!-\!\!* O_k$ or $O_j \leftarrow\!\!* O_k$ as $O_j \bullet\!\!-\!\!* O_k$.*

Proof. For R6a, a filled arrowhead or square at O_j would violate Lemma 11. For R6b and R6c, an unfilled arrowhead at O_j would also violate Lemma 11. For R6d, an unfilled arrowhead at O_j cannot exist because $O_i, O_j \in \boldsymbol{An}(\boldsymbol{S}^t)$ at some $t \in S(f(T))$ by the acyclicity of the CMJ-DAG at any time point.

□

Lemma 17. *The following variations of FCI's R7 are sound under mixture faithfulness, d-separation faithfulness with respect to* $\mathbb{P}_{\boldsymbol{X}^t}, \forall t \in S(f(T))$ *as well as parameter faithfulness:*

1. *R7a: If (1)* $O_i \multimap\!\!* O_j \circ\!\!\multimap\!\!* O_k$ *or* $O_i \multimap\!\!* O_j \bullet\!\!\multimap\!\!* O_k$*, (2)* $O_i \perp\!\!\!\perp O_k | (\boldsymbol{W}, \boldsymbol{S})$ *with* $O_j \in \boldsymbol{W}$ *and minimal independence set* $\boldsymbol{W}$*, and* $f(O_i, O_k | \boldsymbol{B}, \boldsymbol{S})$ *is stationary for all* $\boldsymbol{B} \subseteq (\boldsymbol{W} \setminus O_j)$*, then orient* $O_j \circ\!\!\multimap\!\!* O_k$ *or* $O_j \bullet\!\!\multimap\!\!* O_k$ *as* $O_j \multimap\!\!* O_k$*.*
2. *R7b: If (1)* $O_i \bullet\!\!\multimap\!\!* O_j \circ\!\!\multimap\!\!* O_k$*,* $O_i \blacksquare\!\!\multimap\!\!* O_j \circ\!\!\multimap\!\!* O_k$*,* $O_i \bullet\!\!\multimap\!\!* O_j \blacktriangleleft\!\!\multimap\!\!* O_k$ *or* $O_i \blacksquare\!\!\multimap\!\!* O_j \blacktriangleleft\!\!\multimap\!\!* O_k$*, (2)* $O_i \perp\!\!\!\perp O_k | (\boldsymbol{W}, \boldsymbol{S})$ *with* $O_j \in \boldsymbol{W}$ *and minimal independence set* $\boldsymbol{W}$*, and* $f(O_i, O_k | \boldsymbol{B}, \boldsymbol{S})$ *is stationary for all* $\boldsymbol{B} \subseteq (\boldsymbol{W} \setminus O_j)$*, then orient* $O_j \circ\!\!\multimap\!\!* O_k$ *or* $O_j \blacktriangleleft\!\!\multimap\!\!* O_k$ *as* $O_j \bullet\!\!\multimap\!\!* O_k$*.*
3. *R7c: If (1)* $O_i \multimap\!\!* O_j \circ\!\!\multimap\!\!* O_k$ *or* $O_i \multimap\!\!* O_j \blacktriangleleft\!\!\multimap\!\!* O_k$*, (2)* $O_i \perp\!\!\!\perp O_k | (\boldsymbol{W}, \boldsymbol{S})$ *with* $O_j \in \boldsymbol{W}$ *and minimal independence set* $\boldsymbol{W}$*, and* $f(O_i, O_k | \boldsymbol{B}, \boldsymbol{S})$ *is non-stationary for some* $\boldsymbol{B} \subseteq (\boldsymbol{W} \setminus O_j)$*, then orient* $O_j \circ\!\!\multimap\!\!* O_k$ *or* $O_j \blacktriangleleft\!\!\multimap\!\!* O_k$ *as* $O_j \bullet\!\!\multimap\!\!* O_k$*.*

Proof. R7a: Suppose to the contrary that we have a filled arrowhead at O_j. Then, we must have $O_i \!-\!\!\bullet O_j$ from Lemma 8 due to the minimal separating set; this however contradicts Lemma 11.

R7b: Suppose to the contrary that we have an unfilled arrowhead at O_j. Then, we must have $O_i \bullet\!\!- O_j$ from Lemma 8 due to the minimal separating set; this however contradicts Lemma 11.

R7c: Suppose to the contrary that we have an unfilled arrowhead at O_j. Then, we must have $O_i \!-\!\!\bullet O_j$ from Lemma 8 due to the minimal separating set; this however contradicts Lemma 11.

□

Lemma 18. *The following variations of FCI's R8 are sound:*

1. *R8a: If (1)* $O_i \!-\!\!* O_j \!-\!\!* O_k$*, and (2)* $O_i \circ\!\!\multimap\!\!* O_k$ *or* $O_i \bullet\!\!\multimap\!\!* O_k$*, then orient* $O_i \circ\!\!\multimap\!\!* O_k$ *or* $O_i \bullet\!\!\multimap\!\!* O_k$ *as* $O_i \multimap\!\!* O_k$*.*
2. *R8b: If (1)* $O_i \bullet\!\!\multimap\!\!* O_j \!-\!\!* O_k$ *or* $O_i \blacksquare\!\!\multimap\!\!* O_j \!-\!\!* O_k$*, and (2)* $O_i \circ\!\!\multimap\!\!* O_k$ *or* $O_i \blacktriangleleft\!\!\multimap\!\!* O_k$*, then orient* $O_i \circ\!\!\multimap\!\!* O_k$ *or* $O_i \blacktriangleleft\!\!\multimap\!\!* O_k$ *as* $O_i \bullet\!\!\multimap\!\!* O_k$*.*
3. *R8c: If (1)* $O_i \!-\!\!* O_j \bullet\!\!\multimap\!\!* O_k$ *or* $O_i \!-\!\!* O_j \blacksquare\!\!\multimap\!\!* O_k$*, and (2)* $O_i \circ\!\!\multimap\!\!* O_k$ *or* $O_i \blacktriangleleft\!\!\multimap\!\!* O_k$*, then orient* $O_i \circ\!\!\multimap\!\!* O_k$ *or* $O_i \blacktriangleleft\!\!\multimap\!\!* O_k$ *as* $O_i \bullet\!\!\multimap\!\!* O_k$*.*

4. *R8d: If (1)* $O_i \bullet\!\!-\!\!* O_j \bullet\!\!-\!\!* O_k$, $O_i \blacksquare\!\!-\!\!* O_j \bullet\!\!-\!\!* O_k$, $O_i \bullet\!\!-\!\!* O_j \blacksquare\!\!-\!\!* O_k$ *or* $O_i \blacksquare\!\!-\!\!* O_j \blacksquare\!\!-\!\!* O_k$ *is contemporaneous according to* $O_i -\!\!* O_j -\!\!* O_k$*, and (2)* $O_i \circ\!\!-\!\!* O_k$ *or* $O_i \blacktriangleleft\!\!-\!\!* O_k$*, then orient* $O_i \circ\!\!-\!\!* O_k$ *or* $O_i \blacktriangleleft\!\!-\!\!* O_k$ *as* $O_i \bullet\!\!-\!\!* O_k$.

Proof. All rules follow due to transitivity of the tail.

□

5.3 PSEUDOCODE

I now formally introduce the F^{2}CI algorithm as summarized in pseudocode in Algorithm 6. Like FCI and RFCI, the algorithm proceeds in three main steps: skeleton discovery, v-structure orientation and orientation rule application.

The skeleton discovery procedure of F^{2}CI remains unchanged from that of FCI (Algorithms 1, 2, 3). The v-structure discovery procedure of F^{2}CI mimics that of Algorithm 4 in RFCI, but it requires some additional queries to the CMM oracle in lines 5, 8, and 12 of Algorithm 7 according to Lemma 6. Algorithm 7 also utilizes a time indexing cell $\mathcal{K}$ in lines 2, 10 and 14 to help keep track of contemporaneous endpoints as required by the orientation rules. Here, we may determine that two arbitrary endpoints are contemporaneous, say at O_j for the edge $O_i *\!\!-\!\!* O_j$ and at O_k for the edge $O_k *\!\!-\!\!* O_l$, if we have at least one overlapping value in both $\mathcal{K}_{ij}$ as well as $\mathcal{K}_{kl}$. For example, if $\mathcal{K}_{ij} = \{1, 2\}$ and $\mathcal{K}_{kl} = \{2, 3\}$, then we have the overlapping value 2.

The orientation rule application procedure introduces the most changes compared to the previous steps. First notice that F^{2}CI orients as many edges as possible when applying each orientation rule in line 4 of Algorithm 9. The algorithm must therefore detect all edges which satisfy the sufficient conditions of a rule when applying the rule; this process ensures that F^{2}CI includes all possible time points in $\mathcal{K}$. Second, F^{2}CI calls Algorithm 8 when applying an orientation rule to ensure that it records all time points in $\mathcal{K}$ in lines 4 and 7 of Algorithm 8. Finally, F^{2}CI checks whether any newly added time points in $\mathcal{K}$ result only from non-stationary densities in order to terminate the loop in line 9 of Algorithm 9 due to the following proposition:

Proposition 6. *If Algorithm 9 terminates, then there does exist a sequence of the 9 orientation rules that can introduce a new endpoint in* $\mathbb{G}^M$.

Proof. For proof by contradiction, suppose there exists an arbitrary endpoint η in $\mathbb{G}^M$ which can be oriented by a sequence of the 9 orientation rules. First assume that η is filled. Then, η cannot be contemporaneous with another endpoint because (1) we have $\max \mathcal{K}_{old} < \min\left(\mathcal{K} \setminus \mathcal{K}_{old}\right)$ and (2) Algorithm 9 applies all orientation rules exhaustively. Thus, η must be non-contemporaneous with all other endpoints. But then we must have $\mathbb{G}^M_{old} \neq \mathbb{G}^M$ because Algorithm 9 applies all orientation rules exhaustively. Now suppose that the endpoint is unfilled. But again, if a new unfilled endpoint is added, then we must have $\mathbb{G}^M_{old} \neq \mathbb{G}^M$ because Algorithm 9 applies all orientation rules exhaustively. □

The termination criterion of the loop therefore differs from that of FCI or RFCI which only checks whether $\mathbb{G}^M$ remains unchanged. F^{2}CI may thus continue to apply the orientation rules even if $\mathbb{G}^M$ remains unchanged in an effort to record all necessary time points in $\mathcal{K}$.

Data: CI oracle, CMM oracle
Result: $\mathbb{G}^M$

1 Run FCI's skeleton discovery procedure using Algorithms 1, 2, 3
2 Orient v-structures using Algorithm 7
3 Apply orientation rules using Algorithm 9

Algorithm 6: F^{2}CI

5.4 SUMMARY OF THE OUTPUT

I say that a graph $\mathbb{G}^M$ is an F^{2}CI Partial Time-dependent Ancestral Graph (F^{2}CI-PTAG) that represents the CMJ-DAG $\mathbb{G}$ on $S(f(T))$ if:

1. The absence of an edge between two vertices O_i and O_j in $\mathbb{G}^M$ implies that there exists a subset $\boldsymbol{W} \subseteq \boldsymbol{O} \setminus \{O_i, O_j\}$ such that $O_i \perp\!\!\!\perp O_j | (\boldsymbol{W} \cup \boldsymbol{S})$ in $\mathbb{P}_{\boldsymbol{X}^m}$.
2. The presence of an edge between two vertices O_i and O_j in $\mathbb{G}^M$ implies that $O_i \not\perp\!\!\!\perp O_j | (\boldsymbol{W} \cup \boldsymbol{S})$ in $\mathbb{P}_{\boldsymbol{X}^m}$ for all subsets $\boldsymbol{W} \subseteq \widetilde{\boldsymbol{PDS}}(O_i) \setminus O_j$ and for all $\boldsymbol{W} \subseteq \widetilde{\boldsymbol{PDS}}(O_j) \setminus O_i$.

Data: Skeleton $\mathbb{G}^M$, sepset, $\mathcal{M}$
Result: $\mathbb{G}^M$, $\mathcal{K}$

Run lines 1-19 in Algorithm 4
Let $\mathcal{K}$ denote an empty cell
forall *elements* $\langle O_i, O_k, O_j \rangle$ *of* $\mathcal{L}$ **do**
  if $O_k \notin \text{sepset}(O_i, O_j)$ *and both edges* $O_i \ast\!\!-\!\!\ast O_k$ *and* $O_k \ast\!\!-\!\!\ast O_j$ *are present* **then**
    if *both* $f(O_i, O_k|\boldsymbol{A})$ *and* $f(O_j, O_k|\boldsymbol{A})$ *are stationary* **then**
      Orient $\langle O_i, O_k, O_j \rangle$ as $O_i \ast\!\!\rightarrow O_k \leftarrow\!\!\ast O_j$
    end
    else if $f(O_i, O_k|\boldsymbol{A})$ *is stationary and at least one member of* $\{f(O_k|\boldsymbol{A}), f(O_j|\boldsymbol{A})\}$ *is stationary* **then**
      Orient $\langle O_i, O_k, O_j \rangle$ as $O_i \ast\!\!\rightarrow O_k \blacktriangleleft\!\!\ast O_j$
      Add $1 + \max \mathcal{K}$ to $\mathcal{K}_{ik}$ and $\mathcal{K}_{jk}$
    end
    else if $f(O_j, O_k|\boldsymbol{A})$ *is stationary and at least one member of* $\{f(O_k|\boldsymbol{A}), f(O_i|\boldsymbol{A})\}$ *is stationary* **then**
      Orient $\langle O_i, O_k, O_j \rangle$ as $O_i \ast\!\!\rightarrow O_k \blacktriangleleft\!\!\ast O_j$
      Add $1 + \max \mathcal{K}$ to $\mathcal{K}_{ik}$ and $\mathcal{K}_{jk}$
    end
  end
end

Algorithm 7: V-Structure Discovery for F^{2}CI

Data: r,$\mathbb{G}^M$, $\mathcal{K}$
Result: $\mathbb{G}^M$, $\mathcal{K}$

if *the sufficient conditions of Rule r hold* **then**
  Fire Rule r and accordingly modify $\mathbb{G}^M$
  if *(1) a filled endpoint was just oriented by Rule r at say O_j on the edge $O_i \ast\!\!-\!\!\ast O_j$, (2) all filled edges in the sufficient conditions of Rule r contain overlapping time points $\boldsymbol{K}$ in $\mathcal{K}$, and (3) every density checked by $\mathcal{M}$ is stationary* **then**
    $\mathcal{K}_{ij} \leftarrow \mathcal{K}_{ij} \cup \boldsymbol{K}$
  end
  else if *(1) a filled endpoint was just oriented by Rule r at say O_j on the edge $O_i \ast\!\!-\!\!\ast O_j$, and (2) $\exists$ a density checked by $\mathcal{M}$ that is non-stationary* **then**
    $\mathcal{K}_{ij} \leftarrow \{\mathcal{K}_{ij}, 1 + \max \mathcal{K}\}$
  end
end

Algorithm 8: Rule Application for F^{2}CI

Data: Rule r, $\mathbb{G}^M$, $\mathcal{K}$
Result: $\mathbb{G}^M$, $\mathcal{K}$

repeat
  $\mathbb{G}^M_{old} \leftarrow \mathbb{G}^M$
  for *Rule r in R1a-R10e* **do**
    Orient as many edges as possible using Rule r with Algorithm 8
  end
  if $\mathbb{G}^M_{old} \neq \mathbb{G}^M$ **then**
    $\mathcal{K}_{old} \leftarrow \mathcal{K}$
  end
until $\max \mathcal{K}_{old} < \min\left(\mathcal{K} \setminus \mathcal{K}_{old}\right)$ *and* $\mathbb{G}^M_{old} = \mathbb{G}^M$;

Algorithm 9: Orientation Rules for F^{2}CI

3. If we have the *unfilled arrowhead* $O_i * \!\!\rightarrow O_j$, then $O_j^t \notin \boldsymbol{An}(O_i^t \cup \boldsymbol{S}^t)$ for all $t \in S(f(T))$.

4. If we have the *unfilled tail* $O_i * \!\!- O_j$, then $O_j^t \in \boldsymbol{An}(O_i^t \cup \boldsymbol{S}^t)$ for all $t \in S(f(T))$.

5. If we have the *square* $O_i * \!\!-\!\!\blacksquare O_j$, then $O_j^t \in \boldsymbol{An}(O_i^t \cup \boldsymbol{S}^t)$ for some $t \in S(f(T))$ and $O_j^t \notin \boldsymbol{An}(O_i^t \cup \boldsymbol{S}^t)$ for some other $t \in S(f(T))$.

6. If we have the *filled arrowhead* $O_i * \!\!-\!\!\blacktriangleright O_j$, then we either have $O_i * \!\!\rightarrow O_j$ or $O_i * \!\!-\!\!\blacksquare O_j$.

7. If we have the *filled tail* $O_i * \!\!-\!\!\bullet O_j$, then we either have $O_i * \!\!- O_j$ or $O_i * \!\!-\!\!\blacksquare O_j$.

8. If we have the *circle* $O_i * \!\!-\!\!\circ O_j$, then we either have $O_i * \!\!\rightarrow O_j$, $O_i * \!\!- O_j$, or $O_i * \!\!-\!\!\blacksquare O_j$.

I also propose RF^{2}CI which remains identical to F^{2}CI except we just perform PC skeleton discovery during the skeleton discovery phase. Now a graph $\mathbb{G}^M$ is an RF^{2}CI-PTAG that represents the CMJ-DAG $\mathbb{G}$ on $S(f(T))$ if the aforementioned conditions 1 and 3-8 hold, but we have a modified condition 2:

2. The presence of an edge between two vertices O_i and O_j in $\mathcal{G}$ implies that $O_i \perp\!\!\!\perp O_j | (\boldsymbol{W} \cup \boldsymbol{S})$ in $\mathbb{P}^m$ for all subsets for all subsets $\boldsymbol{W} \subseteq \widetilde{\boldsymbol{Adj}}(O_i) \setminus O_j$ and for all subsets $\boldsymbol{W} \subseteq \widetilde{\boldsymbol{Adj}}(O_j) \setminus O_i$. Here, $\widetilde{\boldsymbol{Adj}}(O_j)$ refers those variables adjacent to O_j after running PC's skeleton discovery procedure.

We conclude this subsection with the following result:

Theorem 2. *Assume mixture faithfulness, d-separation faithfulness with respect to* $\mathbb{P}_{\boldsymbol{X}^t}, \forall t \in S(f(T))$ *and parameter faithfulness. Then, the output of F^{2}CI is an F^{2}CI-PTAG, and the output of RF^{2}CI is an RF^{2}CI-PTAG.*

Proof. F^{2}CI discovers $\widetilde{\boldsymbol{PDS}}(O_i)$ for each $O_i \in \boldsymbol{O}$ by Lemma 5. By a similar argument, RF^{2}CI discovers $\widetilde{\boldsymbol{Adj}}(O_i)$, a superset of $\boldsymbol{Adj}(O_i)$, for each $O_i \in \boldsymbol{O}$. Soundness of the arrowheads in v-structure discovery follows by Lemma 6. Soundness of the other endpoints follows by the soundness of the orientation rules in Lemmas 7, 8 and 11-18. □

5.5 IMPLEMENTATION

I implemented F^{2}CI for the linear Gaussian case due to modern limitations of CMM. According to the previous chapter, the proposed algorithm requires a sound method for assessing whether the number of components in any conditional mixture model does or does not exceed one; we cannot just fit a conditional mixture model with a known number of components. In practice, investigators usually determine the number of components by either using an information criterion score such as BIC or AIC or by performing parametric bootstrap of the likelihood ratio statistic for testing:

$$\begin{aligned} H_0 &: m = m_0, \\ H_1 &: m = m_0 + 1, \end{aligned} \tag{5.1}$$

for some positive integer m_0 [McLachlan, 1987].[2] However, methods for determining the number of components quickly run into theoretical or practical issues once one moves away from parametric and/or linear models. For example, non-parametric methods which allow the user to determine the number of components also often impose a conditional independence assumption which prevents their usefulness in the proposed setting [Allman et al., 2009, Sgouritsa et al., 2013]. Note that a semi-parametric method with automatic model selection currently does exist, but the method does not scale well beyond a univariate conditioning set [Huang et al., 2013]. I suspect however that these limitations will likely disappear in the future as investigators develop better performing and more general semi-parametric methods for conditional mixture modeling.

Although I focus on the linear Gaussian case in the implementation, F^{2}CI is in no way restricted to this situation in theory provided that the user has access to a more general CMM method which can automatically determine whether or not the number of components exceeds one. I have in fact designed the algorithm so that, if the user acquires such a method, then the user can simply use his or her CMM method in F^{2}CI without sacrificing the algorithm's soundness.

[2] Methods for parametric bootstrapping of the likelihood ratio test statistics sequentially test $m = m_0$ versus $m = m_0 + 1$ for $m_0 = 1, 2, \ldots$ The methods then terminate after the bootstrapped p-value for one of these tests exceeds a specified significance level (typically 0.05).

5.5.1 Finite Mixtures of Multiple Response Linear Regressions

I now extend CMM in the Gaussian case to multiple responses. Consider the following Gaussian model:

$$\boldsymbol{Y} = \boldsymbol{X}^T\beta + \boldsymbol{\varepsilon}, \tag{5.2}$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$.

Now assume Σ is non-singular and known. Then we can write the log-likelihood for β conditional on $\boldsymbol{X}$ up to a constant not depending on β as follows:

$$L(\beta) = -\frac{1}{2}\mathrm{tr}\big[(\underline{\boldsymbol{Y}} - \underline{\boldsymbol{X}}\beta)C(\underline{\boldsymbol{Y}} - \underline{\boldsymbol{X}}\beta)^T\big], \tag{5.3}$$

where $C = \Sigma^{-1}$ and $\underline{\boldsymbol{Y}}$ is a matrix with rows corresponding to samples. I can also write the partial derivative of 5.3 with respect to β as follows:

$$\frac{\partial L(\beta)}{\partial \beta} = \underline{\boldsymbol{X}}^T\underline{\boldsymbol{Y}}C^T - \underline{\boldsymbol{X}}^T\underline{\boldsymbol{X}}\beta C^T. \tag{5.4}$$

We obtain the following maximum likelihood estimate $\widehat{\beta}$ after setting the derivative equal to zero and solving for β:

$$\widehat{\beta} = (\underline{\boldsymbol{X}}^T\underline{\boldsymbol{X}})^{-1}\underline{\boldsymbol{X}}^T\underline{\boldsymbol{Y}}, \tag{5.5}$$

Notice that the maximum likelihood estimate does not depend on Σ. As a result, we can obtain $\widehat{\beta}$ by simply performing multiple univariate least square regressions.

We can similarly consider a weighted log-likelihood:

$$L(\beta) = -\frac{1}{2}\mathrm{tr}\big[W(\underline{\boldsymbol{Y}} - \underline{\boldsymbol{X}}\beta)C(\underline{\boldsymbol{Y}} - \underline{\boldsymbol{X}}\beta)^T)\big], \tag{5.6}$$

where W is a diagonal matrix of weights. Again, setting the partial derivative with respect to β to zero and then solving for β, we obtain the following maximum likelihood estimate:

$$\widehat{\beta} = (\underline{\boldsymbol{X}}^T W\underline{\boldsymbol{X}})^{-1}\underline{\boldsymbol{X}}^T W\underline{\boldsymbol{Y}}. \tag{5.7}$$

Now consider the following conditional mixture density:

$$f(\boldsymbol{Y}|\boldsymbol{X}) = \sum_{j=1}^{m} \lambda_j \mathcal{N}(\boldsymbol{Y}|\boldsymbol{X}^T\beta_j, \sigma_j^2). \tag{5.8}$$

We can use the EM algorithm to find a local maximum of the expected multivariate likelihood. The E-step admits a simple modification with a multivariate response:

$$\phi_{ij}^{(t)} = \frac{\lambda_j^{(t)} \zeta(\boldsymbol{y}_i | \boldsymbol{x}_i^T \beta_j^{(t)}, \sigma_j^{2,(t)})}{\sum_{j'=1}^{m} \lambda_{j'}^{(t)} \zeta(\boldsymbol{y}_i | \boldsymbol{x}_i^T \beta_{j'}^{(t)}, \sigma_{j'}^{2,(t)})}. \tag{5.9}$$

Finally, the M-step updates of the β and σ parameters are given by the aforementioned weighted least squares:

$$\begin{aligned} \beta_j^{(t+1)} &= (\underline{\boldsymbol{X}}^T W_j^{(t)} \underline{\boldsymbol{X}})^{-1} \underline{\boldsymbol{X}}^T W_j^{(t)} \underline{\boldsymbol{Y}}, \\ \sigma_j^{2(t+1)} &= \frac{\left\| \sqrt{W_j^{(t)}} (\underline{\boldsymbol{Y}} - \underline{\boldsymbol{X}}^T \beta_j^{(t+1)}) \right\|^2}{\mathrm{tr}(W_j^{(t)})}. \end{aligned} \tag{5.10}$$

6.0 EVALUATION

I now describe the evaluation procedure on synthetic and then real data.

6.1 SYNTHETIC DATA

6.1.1 Algorithms

I compared the following four algorithms: F^{2}CI, RF^{2}CI, FCI and RFCI. I performed causal discovery with the non-parametric KCI test[1] [Zhang et al., 2011] for sample sizes up to and including 500; otherwise, I used a faster non-parametric CI test called RCoT [Strobl et al., 2017]. Finally, I utilized the EM algorithms described in Sections 3.10.2 and 5.5.1 for fitting finite mixtures of linear regressions. All experiments were run on a laptop with 2.6 GHz of CPU and 16 GB of RAM.

6.1.2 Metrics

I will consider CMJs with and without non-stationary feedback. Recall that the usual acyclic causal DAG is a stationary CMJ without feedback. Now, by design, the oracle versions of F^{2}CI and FCI as well as the oracle versions of RF^{2}CI and RFCI give identical results in this usual acyclic scenario. I can therefore evaluate the four algorithms in a straightforward fashion using the structural Hamming distance (SHD) to their oracle graphs for this case [Tsamardinos et al., 2006].

[1] Fisher's z-test is not enough even with a linear Gaussian CMJ due to the potential mixing of the Gaussians.

Observe however that we cannot compute SHD when non-stationarity and/or feedback loops exist because the oracle graphs are not identical between F^{2}CI and FCI as well as between RF^{2}CI and RFCI in this situation. We therefore require a more sophisticated approach.

I choose to use informedness, markedness and Matthew's correlation to evaluate the algorithms when non-stationarity and/or feedback loops exist [Powers]:

$$\begin{aligned}
\text{Informedness} &= \frac{TP}{TP+FN} + \frac{TN}{TN+FP} - 1, \\
\text{Markedness} &= \frac{TP}{TP+FP} + \frac{TN}{TN+FN} - 1, \\
\text{Correlation} &= \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \\
&= \pm\sqrt{\text{Informedness} * \text{Markedness}}.
\end{aligned} \tag{6.1}$$

The above formulations of informedness and markedness are equivalent to those in Equation 3.22. We can view informedness as a bias and prevalence corrected recall measure, markedness as a a bias and prevalence corrected precision measure, and (Matthew's) correlation as a resemblance to the geometric mean between informedness and markedness.

I now must define a true positive, false positive, true negative and false negative in order to use the informedness and markedness metrics. Note that there is no one way to define these quantities. For this thesis, let a positive denote a non-circle endpoint while a negative denote a circle endpoint in the output of the algorithm. I define a true positive as a correctly determined non-circle endpoint; i.e., the ancestral or non-ancestral relation in the output holds in the ground-truth CMJ. A false positive is an incorrectly determined non-circle endpoint. A true negative is a correctly determined circle endpoint; i.e., a circle endpoint which exists on an edge as any endpoint type in the output of the skeleton discovery phase of the algorithm run with oracle information. A false negative is an incorrectly determined circle endpoint.

The above definitions of a true positive, false positive, true negative and false negative have two advantages. First, the definition of a positive allows us to directly compare filled and unfilled non-circle endpoints. We can therefore directly compare F^{2}CI with FCI as well as RF^{2}CI with RFCI. Second, recall that we have identical skeleton discovery phases in F^{2}CI

and FCI as well as in RF^{2}CI and RFCI. Orienting an increasing number of circle endpoints will thus decrease the number of circle endpoints to the same degree in F^{2}CI and FCI as well as in RF^{2}CI and RFCI for the definition of a negative. Now observe that other definitions of a positive and a negative based on (1) ancestral/non-ancestral relations, or (2) separating out all endpoint types, do not inherit both of the aforementioned advantages.

Next recall that the oracle graph adjacencies can appear due to the existence of adjacencies in the CMJ-DAG but also due to the existence of inducing paths and/or parameter dependence. I therefore created a third set of metrics only depending on the ground truth CMJ-DAG rather than the oracle, where we kept the same positives but changed the negatives to CMJ-DAG adjacencies rather than oracle graph adjacencies.

6.1.3 Data Generation

I used the following procedure in [Colombo et al., 2012] to generate 100 different Gaussian DAGs with an expected neighborhood size of $\mathbb{E}(N) = 2$ and $p = 20$ vertices. First, I generated a random adjacency matrix $\mathcal{A}$ with independent realizations of Bernoulli($\mathbb{E}(N)/(p-1)$) random variables in the lower triangle of the matrix and zeroes in the remaining entries. Next, I replaced the ones in $\mathcal{A}$ by independent realizations of a Uniform($[-1, -0.1] \cup [0.1, 1]$) random variable. We can interpret a nonzero entry $\mathcal{A}_{ij}$ as an edge from X_i to X_j with coefficient $\mathcal{A}_{ij}$ in the following linear model:

$$\begin{aligned} X_1 &= \varepsilon_1, \\ X_i &= \sum_{r=1}^{p-1} \mathcal{A}_{ir} X_r + \varepsilon_i, \end{aligned} \tag{6.2}$$

for $i = 2, \ldots, p$ where $\varepsilon_1, ..., \varepsilon_p$ are mutually independent $\mathcal{N}(0, 1)$ random variables. I finally introduced non-zero means μ by adding p independent realizations of a $\mathcal{N}(0, 4)$ random variable to $\boldsymbol{X}$. The variables $X_1, \ldots, X_p$ then have a multivariate Gaussian distribution with mean vector μ and covariance matrix $\Sigma = (\mathbb{I} - \mathcal{A})^{-1}(\mathbb{I} - \mathcal{A})^{-T}$, where $\mathbb{I}$ is the $p \times p$ identity matrix.

I converted the 100 DAGs into 100 CMJ-DAGs with latent and selection variables using the following procedure. For each DAG, I first randomly select with replacement a set of

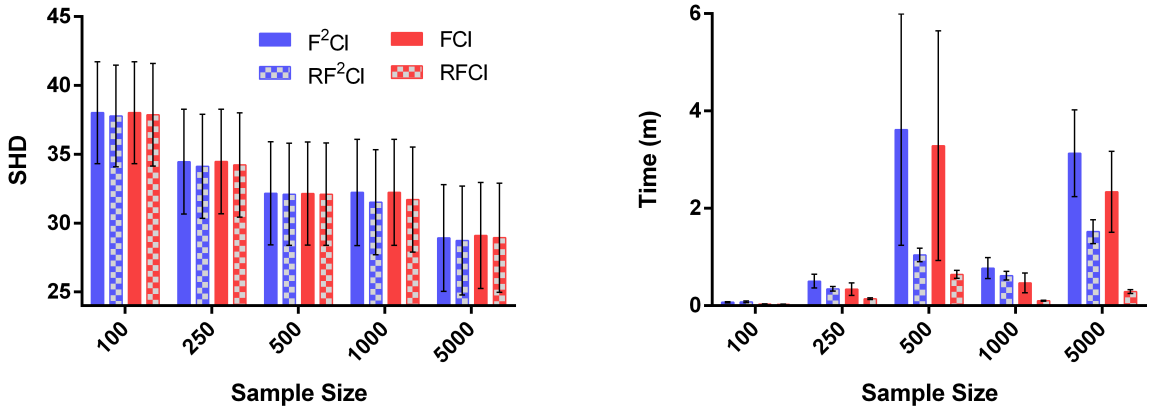
either 0 or 3-6 non-stationary variables from $\boldsymbol{X}$, which I now call $\boldsymbol{D}$. Here, a non-stationary variable denotes a variable that may contain multiple nodes in the CMJ-DAG; as a result, non-stationary variables may introduce non-stationary distributions as well as feedback. I next drew the length of time for each causal relation to any variable in $\boldsymbol{D}$ from $[0.1, 0.5]$. I similarly randomly selected a set of 0-3 latent common causes $\boldsymbol{L}$ without replacement. From the set of $\boldsymbol{X} \setminus \{\boldsymbol{D}, \boldsymbol{L}\}$, I then selected a set of 0-3 selection variables $\boldsymbol{S}$ without replacement.

I sampled each of the resulting CMJ-DAGs as follows. I first sampled each CMJ-DAG uniformly from time point 0 to its ending time point plus 0.5 time points. For each selection variable in $\boldsymbol{S}$, I then eliminated the bottom s percentile of samples, where I drew s according to independent realizations of a Uniform($[0.1, 0.5]$) random variable. I finally eliminated all of the instantiations of the latent variables from the dataset. Ultimately, I created five datasets with sample sizes of 100, 200, 500, 1000 and 5000 for each of the 100 CMJ-DAGs.

6.1.4 Results without Non-Stationarity & Feedback

I first evaluated the algorithms in a CMJ without non-stationary distributions and feedback loops by setting $|\boldsymbol{D}| = 0$; this situation is equivalent to causal discovery with the usual causal DAG. Here, we hope F^{2}CI and FCI as well as RF^{2}CI and RFCI will give near identical results.

I have summarized the results in Figure 6.1 in terms of the SHD to the oracle graphs as well as computation time. None of the pair-wise comparisons of the mean SHDs between F^{2}CI and FCI as well as between RF^{2}CI and RFCI reached statistical significance at any sample size even at an uncorrected threshold of 0.05 using paired t-tests (max absolute t-value: 1.578, min two-sided p-value: 0.118). I therefore conclude that F^{2}CI and FCI have comparable performance in the acyclic case, and likewise for RF^{2}CI and RFCI. In fact, both F^{2}CI and RF^{2}CI achieve this comparable performance under 4 minutes on average across all sample sizes. Recall that the algorithms exhibit a large drop in run-time at a sample size of 1000, because I used the faster CI test called RCoT instead of KCIT for sample sizes above 500.



(a) (b)

Figure 6.1: Results for the acyclic case in terms of SHD to the oracle graphs as well as computation time. Error bars denote 95% confidence intervals of the mean.

6.1.5 Results with Non-Stationarity & Feedback

I have summarized the results for CMJs with non-stationary distributions and/or feedback loops in Figure 6.2. Here, I set $|\boldsymbol{D}|$ to 3, 4, 5 or 6. Notice that F^{2}CI and RF^{2}CI continue to improve across the three metrics of informedness, markedness and Matthew's correlation based on the oracle graph adjacencies with increasing sample sizes whereas FCI and RFCI saturate. I have summarized the t-statistics of paired t-tests across all of the three metrics and all of the sample sizes in Table 6.1. Observe in particular the large magnitudes of t-statistics for all of the three metrics. I therefore conclude that the proposed methods outperform FCI and RFCI by a large margin.

Note that virtually the same results held even when I gave FCI and RFCI the benefit of the doubt by considering a relaxed interpretation of their tail endpoints; here, I count $O_i -\!\!* O_j$ as correct if $O_i^t \in \boldsymbol{An}(O_j^t, \boldsymbol{S}^t)$ at some $t \in S(f(T))$ (as opposed to all $t \in S(f(T))$) for FCI and RFCI only. I maintained the more stringent interpretation of all time points for the unfilled tails in the output of F^{2}CI and RF^{2}CI. I have listed the t-statistics for this case in

Table 6.2; notice that the t-statistic values are only slightly smaller than those of Table 6.1. Finally, F^{2}CI still outperforms FCI and RF^{2}CI still outperforms RFCI using informedness, markedness and correlation based on the CMJ-DAG adjacencies (Figure 6.3). Here, recall that the true positives and false positives remain the same, but the true negatives correspond to circle endpoints which exist on any edge as any endpoint type in the CMJ-DAG; thus a true negative corresponds to a correctly determined adjacency. The false negatives, on the other hand, correspond to incorrectly determined adjacencies, or circle endpoints which do not exist on any edge in the CMJ-DAG.

I have also summarized the computational time of the four algorithms in part (d) of Figure 6.2. In general, F^{2}CI and RF^{2}CI take longer to complete than FCI and RFCI, respectively, due to the additional conditional mixture modeling. Note that the algorithms exhibit a large drop in run time at a sample size of 1000, because we equipped the algorithms with the faster RCoT CI test for sample sizes of 1000 and 5000 (as opposed to the KCIT CI test). However, all algorithms complete within 7 minutes on average across all sample sizes.

6.2 REAL DATA

6.2.1 Algorithms

I again compared F^{2}CI, RF^{2}CI, FCI and RFCI. We equipped the algorithms with the RCoT test as well as the Gaussian regression CMM methods as used in the previous section.

6.2.2 Metrics

We usually do not have access to the ground truth CMJ with real data. We can nevertheless still use some additional information in order to ascertain some of the underlying causal structure. In this thesis, I consider longitudinal datasets that have additional time information. Recall that causal relations cannot occur backwards in time under the CMJ framework; i.e., $O_i^{J_{i,k}}$ cannot be an ancestor of $O_j^{J_{j,l}}$, if $J_{j,l} < J_{i,k}$. I therefore evaluated the four algorithms using real longitudinal datasets by first running the algorithms on the

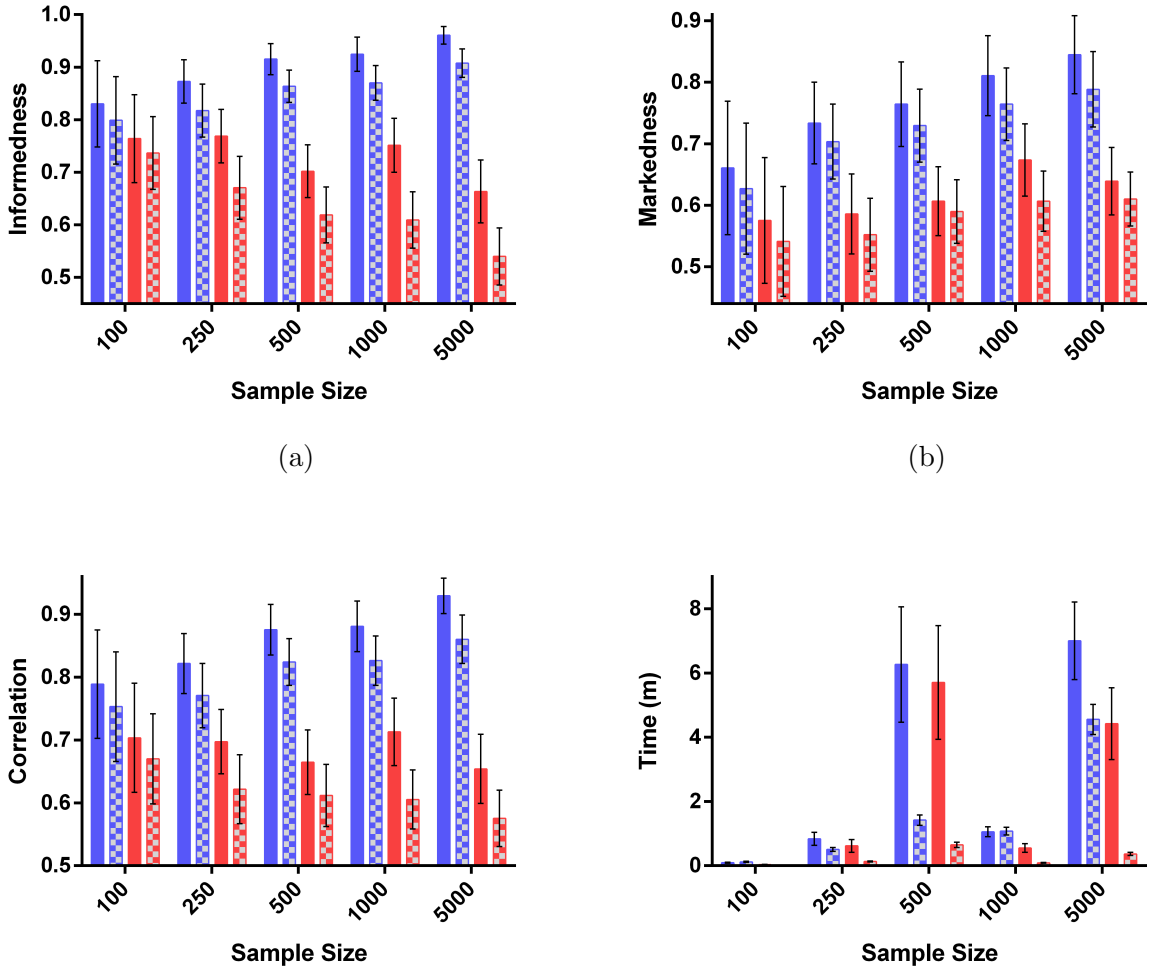


Figure 6.2: Results for the non-stationarity and feedback case in terms of (a) informedness, (b) markedness, (c) Matthew's correlation and (d) computation time. We computed informedness, markedness and Matthew's correlation with the oracle graph adjacencies in this case. Notice that F^{2}CI outperforms FCI by a large margin across the first three metrics, and likewise for RF^{2}CI and RFCI.

longitudinal datasets stripped of time information. Then, I counted the number of errors made by the algorithms by summing over (1) the number unfilled tails, filled tails or squares

	100	250	500	1000	5000
F^{2}CI vs. FCI	*2.727*	4.830	9.684	7.649	10.368
RF^{2}CI vs. RFCI	3.221	7.275	12.298	13.151	17.245

(a)

	100	250	500	1000	5000
F^{2}CI vs. FCI	*2.692*	4.623	5.121	4.175	6.676
RF^{2}CI vs. RFCI	*2.597*	5.408	5.613	5.449	5.664

(b)

	100	250	500	1000	5000
F^{2}CI vs. FCI	*2.853*	4.946	9.446	6.582	10.744
RF^{2}CI vs. RFCI	3.219	6.912	10.168	10.529	13.667

(c)

Table 6.1: T-statistics of (a) informedness, (b) markedness and (c) Matthew's correlation based on the oracle graph adjacencies as a function of sample size. Italicized values did not pass the Bonferroni corrected threshold of 0.05/10.

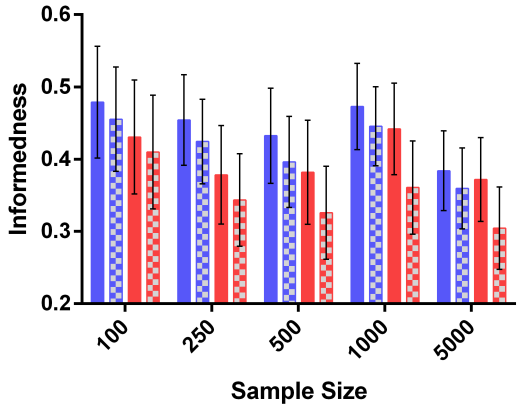
	100	250	500	1000	5000
F^{2}CI vs. FCI	*2.727*	4.677	9.211	7.408	9.630
RF^{2}CI vs. RFCI	3.221	7.103	11.853	12.707	16.668

(a)

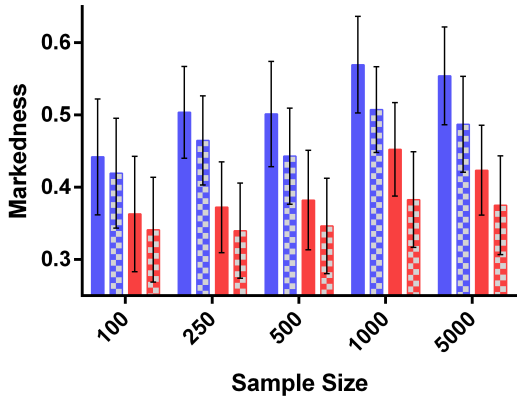
	100	250	500	1000	5000
F^{2}CI vs. FCI	*2.853*	4.777	8.872	6.186	9.890
RF^{2}CI vs. RFCI	3.219	6.680	9.655	10.113	13.202

(b)

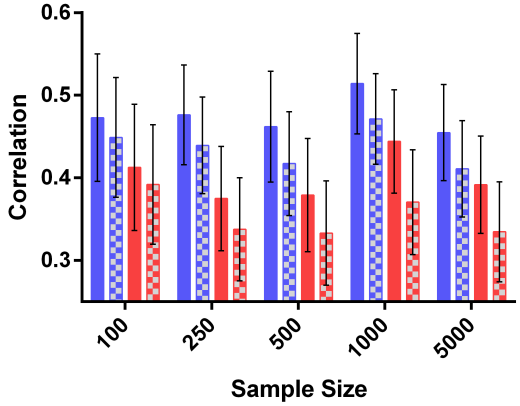
Table 6.2: T-statistics of (a) informedness and (b) Matthew's correlation based on oracle graph adjacencies as a function of sample size using the relaxed tail endpoint interpretation. Markedness values remain identical to those of Table 6.1. Italicized values did not pass the Bonferroni corrected threshold of 0.05/10.



(a)



(b)



(c)

Figure 6.3: Results for the non-stationarity and feedback case in terms of (a) informedness, (b) markedness, (c) Matthew's correlation based on the CMJ-DAG adjacencies. F^{2}CI still outperforms FCI on average, and likewise for RF^{2}CI and RFCI.

from a variable in a later wave to a variable in an earlier wave each with an additional unfilled arrowhead at the variable in the later wave, and (2) the number of unfilled tails from a variable in a later wave to a variable in an earlier wave each with an additional unfilled arrowhead, filled arrowhead or square at the variable in the later wave.

6.2.3 Datasets

I evaluated the algorithms on 200 non-paramateric bootstraps of the following real longitudinal datasets:

1. Continuous clinical data from the Framingham Heart Study [Mahmood et al., 2014]: The Framingham Heart Study is a longitudinal study of the cardiovascular health of residents of Framingham, Massachusetts. I used an abbreviated version of the dataset consisting of 2003 samples and 23 continuous clinical variables across 3 waves after some data cleaning; here, I specifically removed all discrete variables, all variables with more than 1000 missing values and then all samples with any missing values. See Table B1a in Appendix B for a list of all of the final variables.

2. Dahlberg and Johansson's Municipal Expenditure Dataset [Dahlberg and Johansson, 2000]: This dataset lists the expenditures, revenues and grants of 265 Swedish municipalities from 1979 to 1987. The dataset therefore contains a total of 265 samples over $3 * 9 = 27$ variables. See Table B1b in Appendix B for a list of all of the final variables.

Now real data is usually non-Gaussian, so we also subjected both the Framingham and Municipalities datasets to the multivariate Yeo-Johnson transformation to Gaussianity in order to desensitize the Gaussian mixture modeling to deviations from Gaussianity [Yeo, 2000]. We can view the Yeo-Johnson transformation as a more sophisticated log-transform; thus, the Yeo-Johnson transformation does not transform the data to exact Gaussianity except under very special cases. The transformation ultimately helped both F^2CI and RF^2CI orient endpoints; without the transformation, both F^2CI and RF^2CI do not orient any endpoints.

The Yeo-Johnson transformation however was not enough to orient endpoints for the Municipalities dataset due to large deviations from Gaussianity. I therefore also replaced the BIC score with the more conservative Integrated Complete Likelihood (ICL) score[2] during v-structure discovery in order to orient some endpoints in this case.

[2]Recall that the BIC score approximates the integrated likelihood over the observed data. The ICL score on the other hand corresponds to the integrated likelihood over the observed and missing data. Experiments show that the ICL score tends to be more conservative than the BIC score but also more robust to violations of the mixture modeling assumptions [Biernacki et al., 2000].

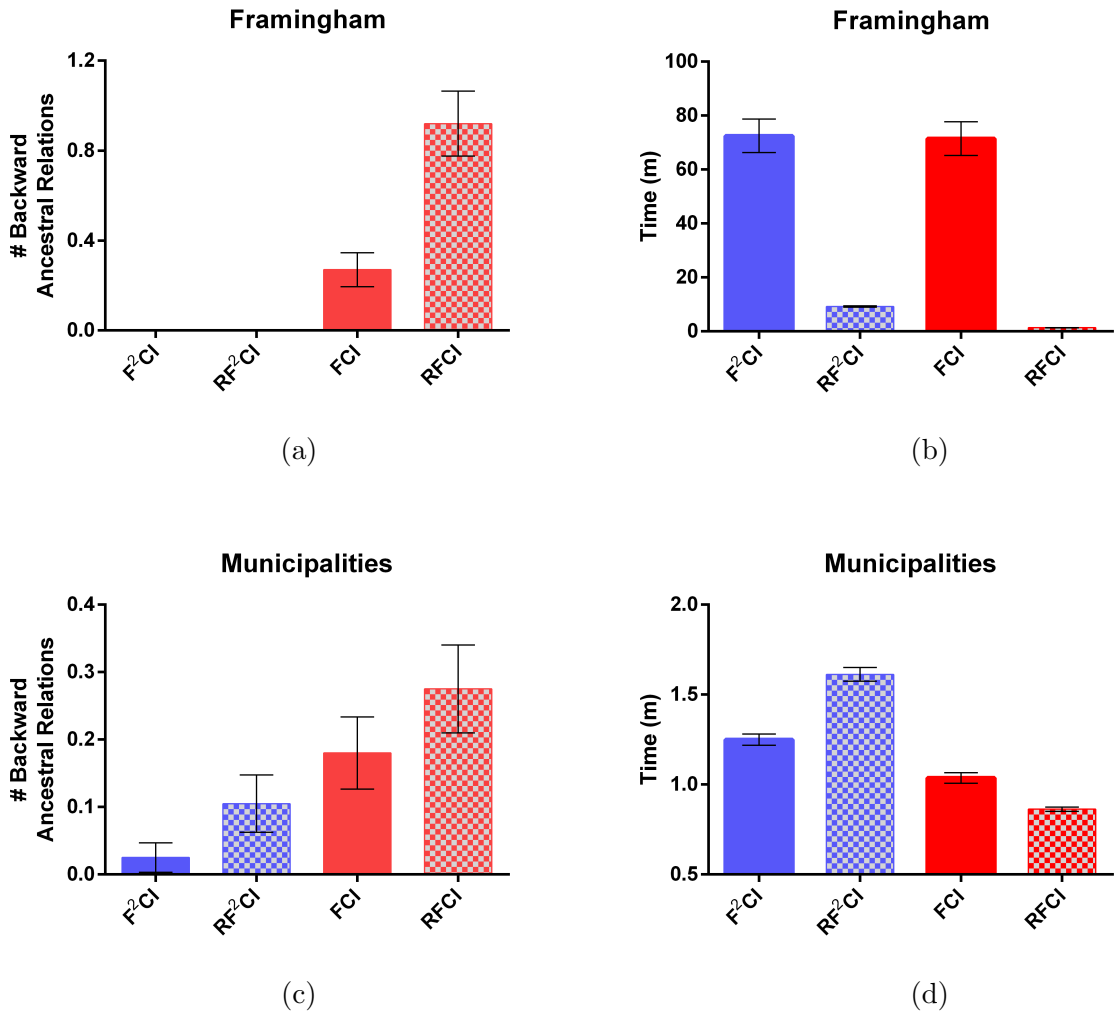


Figure 6.4: The number of ancestral relations directed backwards in time for (a) the Framingham Heart Study and (c) the Swedish Municipalities datasets. F^{2}CI detects less backwards ancestral relations than FCI and likewise for RF^{2}CI and RFCI. Associated timing results also in (b) and (d).

6.2.4 Results

I have summarized the results in Figure 6.4. For the Framingham Heart dataset, I found that F^{2}CI oriented fewer ancestral relations directed backwards in time than FCI under the

Bonferroni level of 0.05/2 (t = -3.019, p = 5.246E-3). A similar result held for RF^{2}CI vs RFCI (t= -6.621, p = 3.244E-10). The results also held even after dividing by the total number of oriented endpoints (F^{2}CI vs FCI: t=-2.896, p = 7.112E-3; RF^{2}CI vs RFCI: t=-12.866, p < 2.200E-16). See Table B2 of Appendix B for a list of all of the inferred endpoints for the Framingham dataset. Finally, F^{2}CI only took 1.030 minutes longer than FCI (95% CI: 0.946, 1.114), while RF^{2}CI took 7.832 minutes longer than RFCI (95% CI: 7.602, 8.062).

The results with the Municipal Expenditure Dataset mimicked the results with the data of the Framingham Heart Study. Again, F^{2}CI and RF^{2}CI oriented fewer ancestral relations directed backwards in time than FCI and RFCI, respectively (F^{2}CI vs FCI: t=-4.932, p = 1.716E-6; RF^{2}CI vs RFCI: t=-7.271, p = 8.012E-12). The results held even after dividing by the total number of oriented endpoints (F^{2}CI vs FCI: t=-4.212, p = 3.968E-5; RF^{2}CI vs RFCI: t=-3.294, p = 1.17E-3). See Table B3 of Appendix B for a list of all of the inferred endpoints for the first 3 waves of this Municipalities dataset. Furthermore, timing results revealed that F^{2}CI only took 0.213 minutes longer than FCI (95% CI: 0.192, 0.233), while RF^{2}CI took 0.750 minutes longer than RFCI (95% CI: 0.716, 0.783). RF^{2}CI also took longer than F^{2}CI in this case, because the graph for RF^{2}CI had many more edges than that of F^{2}CI, so RF^{2}CI had to call the CMM method more times than F^{2}CI.

7.0 CONCLUSION

I now conclude this thesis with a summary in Section 7.1, a discussion of the limitations in Section 7.2, and suggestions for future work in Section 7.3. I then wrap up the thesis with final remarks in Section 7.4.

7.1 SUMMARY

I have developed a framework called the CMJ and an associated algorithm called F^{2}CI which both provide a solution to the problem of causal discovery under non-stationary feedback. The CMJ framework implies that we can view the problem of causal discovery under non-stationary feedback as equivalent to the problem of causal discovery under mixing. I therefore propose an algorithm which uses one type of un-mixing engine called a CMM oracle in order to recover a summary graph of the underlying CMJ. In practice, F^{2}CI outperforms FCI when non-stationarity and/or feedback exists and also performs on par with FCI when they do not exist, as in the usual acyclic case.

7.2 LIMITATIONS

The F^{2}CI algorithm nonetheless carries some limitations due to CMM. First, investigators usually implement the CMM method with the EM algorithm which may not always globally maximize the log-likelihood. The CMM method may therefore not consistently identify

the number of mixture components.[1] Second, CMM methods currently suffer from non-identifiability issues in the non-parametric setting and thus require parametric assumptions. For example, I imposed the Gaussian distribution assumption in the experiments. Extending F^{2}CI to the non-parametric setting may therefore require additional information in order to avoid CMM modeling altogether.

7.3 FUTURE WORK

We may in particular consider utilizing *approximate* time information in future work. Recall that mixture data does not contain time information about the underlying CMJ, so the F^{2}CI algorithm does not use time information. We nevertheless often have access to an approximate time point for each sample in practice (even though we may not have access to the exact time point). For instance, we may consider "years since the diagnosis of mild cognitive impairment" as an approximate time point for Alzheimer's disease, since patients diagnosed with mild cognitive impairment at an earlier date are likely to have more advanced Alzheimer's disease. I believe that the use of approximate time information may allow the development of fully non-parametric causal discovery algorithms that are sound even under non-stationarity and/or feedback.

The importance of the parametric CMM method in the proposed F^{2}CI algorithm nonetheless also implies that we should consider improving CMM modeling by expanding the family of mixing distributions. We can often define such a family over a small class of distributions (e.g., the Gaussians), but we require a more sophisticated family if we want to impose less restrictive assumptions. Here, we must define a family general enough to accommodate most real-world distributions but also stringent enough to allow identifiability of at least some of the mixing parameters; I believe that this is a difficult but interesting open problem for future research in causal discovery.

[1] I nevertheless find that the EM algorithm with the BIC score usually outputs the correct number of components in practice, even when the algorithm does not discover a global optimum.

7.4 FINAL COMMENTS

Note that many other avenues for future work exist, but I encourage researchers to use F^{2}CI as well as its faster sister algorithm RF^{2}CI in the meantime. I have provided an R package containing the algorithms as well as KCIT, RCoT and the mixture of linear regressions methods.[2] I have also designed the algorithms so that investigators can easily substitute in any consistent CI test and/or CMM method suitable for their problem. I ultimately hope that both F^{2}CI and RF^{2}CI will help investigators discover many useful causal relations in their domains of interest.

[2] R package URL: https://github.com/ericstrobl/F2CI

APPENDIX A

SMOKING TOBACCO & LUNG CANCER

I briefly summarize the history of the discovery of the tobacco-lung cancer causal link. Several retrospective observational studies revealed a strong correlation between smoking tobacco and lung cancer beginning in 1940 [Müller, 1940, Schairer and Schöniger, 1944]. In order to eliminate the influence of recall bias, investigators further confirmed the association in prospective cohort studies, even when investigators matched participants by age, sex, occupation and several other traits [Doll and Hill, 2004, Hammond and Horn, 1954]. Investigators also replicated the association between smoking tobacco and lung cancer in multiple twin studies which confirmed that genetics alone could not explain the association between smoking tobacco and lung cancer [Kaprio and Koskenvuo, 1989, 1990, Braun et al., 1994]. Other researchers also identified multiple chemical constitutes of tobacco that were also associated with cancer in human [Roffo, 1939, Rep, 1952]. Next, investigators conducting detailed cellular pathology studies in humans found an association between cigarette smoking and ciliastasis, or the destruction of tiny hair like structures in the airways [Hilding, 1956, Auerback et al., 1957]. Moreover, the cilia were destroyed in precisely those areas where cancers were most likely to develop. Scientists finally found that smearing tobacco on the skins of animals increased tumor development in multiple experiments spanning several animal species [Roffo, 1931, Wynder et al., 1953].

APPENDIX B

REAL DATA VARIABLES & RESULTS

Variable	Waves
Total Cholesterol Level	1,2,3
Age	1,2,3
Systolic Blood Pressure	1,2,3
Diastolic Blood Pressure	1,2,3
Body Mass Index	1,2,3
Heart Rate	1,2,3
Blood Glucose Level	1,2,3
HDL Level	3
LDL Level	3

(a)

Variable	Waves
Expenditures	1-9
Receipts, Taxes & Fees	1-9
Government Grants & Shared Tax Revenues	1-9

(b)

Table B1: Variables and their wave numbers for (a) the Framingham Heart Study and (b) the Municipal Expenditure datasets.

Start	End	Endpoint Type(s)
Heart Rate (1)	Heart Rate (2)	Unfilled & Filled Arrowhead
Heart Rate (3)	Heart Rate (2)	Unfilled & Filled Arrowhead
Heart Rate (1)	Heart Rate (3)	Filled Arrowhead
Glucose (3)	Heart Rate (3)	Filled Arrowhead
Heart Rate (2)	Heart Rate (3)	Filled Arrowhead
Heart Rate (1)	Heart Rate (2)	Filled Arrowhead
Glucose (2)	Heart Rate (2)	Filled Arrowhead
Heart Rate (3)	Heart Rate (2)	Filled Arrowhead
BMI (1)	Total Cholesterol (1)	Filled Arrowhead
BMI (1)	Total Cholesterol (1)	Filled Arrowhead
BMI (2)	Total Cholesterol (1)	Filled Arrowhead
Diastolic Blood Pressure (1)	Diastolic Blood Pressure (2)	Filled Arrowhead
Glucose (2)	Diastolic Blood Pressure (2)	Filled Arrowhead
Glucose (1)	Heart Rate (1)	Filled Arrowhead
Heart Rate (3)	Heart Rate (1)	Filled Arrowhead
BMI (3)	BMI (2)	Filled Tail

Table B2: Edges with non-circle endpoints present in the output of F^2CI for the Framingham Heart Study dataset. The first column denotes the starting vertex and the second column denotes the ending vertex for each edge. Numbers in the parentheses denote the wave number of any given variable. Endpoints in the third column are located at the ending vertex. Each row corresponds to an unique oriented edge present in the output of F^2CI for at least one bootstrap of the Framingham Heart Study dataset.

Start	End	Endpoint Type(s)
Expenditures (1)	Receipts, Taxes & Fees (1)	Unfilled Arrowhead
Receipts, Taxes & Fees (2)	Receipts, Taxes & Fees (1)	Unfilled Arrowhead
Receipts, Taxes & Fees (1)	Receipts, Taxes & Fees (2)	Filled Arrowhead
Receipts, Taxes & Fees (3)	Receipts, Taxes & Fees (2)	Filled Arrowhead
Receipts, Taxes & Fees (1)	Receipts, Taxes & Fees (3)	Filled Arrowhead
Receipts, Taxes & Fees (2)	Receipts, Taxes & Fees (3)	Filled Arrowhead
Expenditures (2)	Expenditures (3)	Filled Arrowhead
Receipts, Taxes & Fees (3)	Expenditures (3)	Filled Arrowhead
Receipts, Taxes & Fees (3)	Receipts, Taxes & Fees (1)	Filled Tail

Table B3: Edges with non-circle endpoints present in the output of F^2CI for the first 3 waves of the Sweden Municipal Expenditure dataset; the full table for all 9 waves contains more than 40 rows. This table otherwise preserves the same format as Table B2. Each row corresponds to an unique oriented edge present in the output of F^2CI for at least one bootstrap of the Sweden Municipal Expenditures dataset.

BIBLIOGRAPHY

Report of progress - technical research department. *Ness Motley Law Firm Documents*, 1952. URL https://www.industrydocumentslibrary.ucsf.edu/tobacco/docs/zkln0042.

E. S. Allman, C. Matias, and J. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132, Dec. 2009. doi: 10.1214/09-AOS689. URL http://projecteuclid.org/euclid.aos/1250515381.

N. Atienza, J. Garca-Heras, J. Muoz-Pichardo, and R. Villa. On the consistency of MLE in finite mixture models of exponential families. *Journal of Statistical Planning and Inference*, 137(2):496 – 505, 2007. ISSN 0378-3758. doi: http://dx.doi.org/10.1016/j.jspi.2005.12.014. URL http://www.sciencedirect.com/science/article/pii/S0378375806000425.

O. Auerback, J. B. Gere, J. B. Forman, T. G. Petrick, H. J. Smolin, G. E. Muehsam, D. Y. Kassouny, and A. P. Stout. Changes in the bronchial epithelium in relation to smoking and cancer of the lung: A report of progress. *New England Journal of Medicine*, 256(3): 97–104, 1957.

T. Benaglia, D. Chauveau, D. R. Hunter, and D. Young. mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1–29, 2009. URL http://www.jstatsoft.org/v32/i06/.

C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 22(7):719–725, July 2000. ISSN 0162-8828. doi: 10.1109/34.865189. URL http://dx.doi.org/10.1109/34.865189.

R. A. Boyles. On the convergence of the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(1):47–50, 1983. ISSN 00359246. URL http://www.jstor.org/stable/2345622.

M. Braun, N. Caporaso, R. Hoover, and W. Page. Genetic component of lung cancer: cohort study of twins. *The Lancet*, 344(8920):440 – 443, 1994. ISSN 0140-6736. doi: http://dx.doi.org/10.1016/S0140-6736(94)91770-1. URL http://www.sciencedirect.com/science/article/pii/S0140673694917701. Originally published as Volume 2, Issue 8920.

D. Colombo, M. Maathius, M. Kalisch, and T. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *Annals of Statistics*, 40(1):294–321, Apr. 2012. doi: 10.1214/11-AOS940. URL http://projecteuclid.org/euclid.aos/1333567191.

P. Dagum, A. Galper, E. Horvitz, and S. U. M. C. S. K. S. Laboratory. *Temporal Probabilistic Reasoning: Dynamic Network Models for Forecasting*. Report (Stanford University. Medical Computer Science. Knowledge Systems Laboratory). Knowledge Systems Laboratory, Medical Computer Science, Stanford University, 1991. URL https://books.google.com/books?id=bOS3PAAACAAJ.

P. Dagum, A. Galper, and E. Horvitz. Dynamic network models for forecasting. In *Proceedings of the Eighth International Conference on Uncertainty in Artificial Intelligence*, UAI'92, pages 41–48, San Francisco, CA, USA, 1992. Morgan Kaufmann Publishers Inc. ISBN 1-55860-258-5. URL http://dl.acm.org/citation.cfm?id=2074540.2074546.

P. Dagum, A. Galper, E. Horvitz, and A. Seiver. Uncertain reasoning and forecasting. *International Journal of Forecasting*, 11:73–87, 1995.

M. Dahlberg and E. Johansson. An examination of the dynamic behaviour of local governments using GMM bootstrapping methods. *Journal of Applied Econometrics*, 15(4): 401–416, 2000. URL http://EconPapers.repec.org/RePEc:jae:japmet:v:15:y:2000:i:4:p:401-416.

D. Dash. Restructing dynamic causal systems in equilibrium. In *Proceedings of the Twenty First Conference on Uncertainty in Artificial Intelligence*, 2005.

R. De Veaux. Mixtures of linear regressions. *Computational Statistics and Data Analysis*, 8: 227–245, 1989.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

R. Doll and A. B. Hill. The mortality of doctors in relation to their smoking habits: a preliminary report. *BMJ*, 328(7455):1529–1533, 2004. ISSN 0959-8138. doi: 10.1136/bmj.328.7455.1529. URL http://www.bmj.com/content/328/7455/1529.

R. J. Evans. Graphs for margins of bayesian networks. *Scandinavian Journal of Statistics*, 43(3):625–648, 2016.

B. S. Everitt and D. J. Hand. *Finite mixture distributions*. Monographs on applied probability and statistics. Chapman and Hall, London, New York, 1981. ISBN 0-412-22420-8. URL http://opac.inria.fr/record=b1090915.

A. Ghassami, S. Salehkaleybar, N. Kiyavash, and K. Zhang. Learning causal structures using regression invariance. *CoRR*, abs/1705.09644, 2017. URL http://arxiv.org/abs/1705.09644.

K. Gopalratnam, H. A. Kautz, and D. S. Weld. Extending continuous time Bayesian networks. In *Proceedings, The Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference, July 9-13, 2005, Pittsburgh, Pennsylvania, USA*, pages 981–986, 2005. URL http://www.aaai.org/Library/AAAI/2005/aaai05-155.php.

E. Hammond and D. Horn. The relationship between human smoking habits and death rates: A follow-up study of 187,766 men. *Journal of the American Medical Association*, 155(15): 1316–1328, 1954. doi: 10.1001/jama.1954.03690330020006. URL +http://dx.doi.org/10.1001/jama.1954.03690330020006.

A. C. Hilding. On cigarette smoking, bronchial carcinoma and ciliary action. I. Smoking habits and measurement of smoke intake. *New England Journal of Medicine*, 254(17): 775–81, 1956.

M. Huang, R. Li, and S. Wang. Nonparametric mixture of regression models. *Journal of the American Statistical Association*, 108(503):929–941, 2013. doi: 10.1080/01621459.2013.772897. URL http://dx.doi.org/10.1080/01621459.2013.772897.

A. Hyttinen, P. O. Hoyer, F. Eberhardt, and M. Järvisalo. Discovering cyclic causal models with latent variables: A general SAT-based procedure. In *Proceedings of the Twenty Ninth Conference on Uncertainty in Artificial Intelligence, UAI 2013, Bellevue, WA, USA, August 11-15, 2013*, 2013. URL https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=2391&proceeding_id=29.

S. P. Iyer, I. Shafran, D. Grayson, K. Gates, J. T. Nigg, and D. A. Fair. Inferring functional connectivity in MRI using Bayesian network structure learning with a modified PC algorithm. *Neuroimage*, 75:165–175, Jul 2013.

J. Kaprio and M. Koskenvuo. Twins, smoking and mortality: A 12-year prospective study of smoking-discordant twin pairs. *Social Science and Medicine*, 29(9):1083–1089, 1989.

J. Kaprio and M. Koskenvuo. Cigarette smoking as a cause of lung cancer and coronary heart disease. a study of smoking-discordant twin pairs. *Acta Geneticae Medicae Et Gemellologiae: Twin Research*, 39(1):2534, 1990. doi: 10.1017/S0001566000005560.

G. Lacerda, P. Spirtes, J. Ramsey, and P. O. Hoyer. Discovering cyclic causal models by independent components analysis. In *Proceedings of the Twenty Fourth Conference in Uncertainty in Artificial Intelligence, Helsinki, Finland, July 9-12, 2008*, pages 366–374, 2008. URL https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=1337&proceeding_id=24.

S. L. Lauritzen and T. S. Richardson. Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):321–348, 2002. ISSN 1467-9868. doi: 10.1111/1467-9868.00340. URL http://dx.doi.org/10.1111/1467-9868.00340.

S. L. Lauritzen, A. P. Dawid, B. N. Larsen, and H. G. Leimer. Independence properties of directed Markov fields. *Networks*, 20(5):491–505, Aug. 1990. doi: 10.1002/net.3230200503. URL http://dx.doi.org/10.1002/net.3230200503.

S. S. Mahmood, D. Levy, R. S. Vasan, and T. J. Wang. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *The Lancet*, 383(9921): 999 – 1008, 2014. ISSN 0140-6736. doi: http://doi.org/10.1016/S0140-6736(13)61752-3. URL http://www.sciencedirect.com/science/article/pii/S0140673613617523.

G. J. McLachlan. On bootstrapping the likelihood ratio test stastistic for the number of components in a normal mixture. *Applied Statistics*, 36(3):318–324, 1987. ISSN 0035-9254. doi: 10.2307/2347790. URL http://dx.doi.org/10.2307/2347790.

C. Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, pages 403–410, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1-55860-385-9. URL http://dl.acm.org/citation.cfm?id=2074158.2074204.

A. Y. Mitrophanov and E. A. Groisman. Positive feedback in cellular control systems. *BioEssays*, 30(6):542–555, 2008. ISSN 1521-1878. doi: 10.1002/bies.20769. URL http://dx.doi.org/10.1002/bies.20769.

F. H. Müller. Tabakmißbrauch und lungencarcinom. *Zeitschrift für Krebsforschung*, 49(1): 57–85, 1940. ISSN 1432-1335. doi: 10.1007/BF01633114. URL http://dx.doi.org/10.1007/BF01633114.

U. Nodelman, C. R. Shelton, and D. Koller. Continuous time Bayesian networks. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, pages 378–387, 2002.

U. Nodelman, C. R. Shelton, and D. Koller. Learning continuous time Bayesian networks. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, UAI'03, pages 451–458, San Francisco, CA, USA, 2003. Morgan Kaufmann Publishers Inc. ISBN 0-127-05664-5. URL http://dl.acm.org/citation.cfm?id=2100584.2100639.

U. Nodelman, C. R. Shelton, and D. Koller. Expectation maximization and complex duration distributions for continuous time Bayesian networks. In *Proceedings of the 21st International Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2005. ISBN 0-9749039-1-4. URL http://dblp.uni-trier.de/db/conf/uai/uai2005.html#NodelmanSK05.

L. E. Ortiz and L. P. Kaelbling. Accelerating EM: An empirical study. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, pages 512–521, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1-55860-614-9. URL http://dl.acm.org/citation.cfm?id=2073796.2073854.

J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 78(5):947–1012, 2016. URL http://arxiv.org/abs/1501.01332.

D. M. W. Powers. Evaluation: From precision, recall and f-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, (1):37–63.

R. E. Quandt and J. B. Ramsey. Estimating mixtures of normal distributions and switching regressions. *Journal of the American Statistical Association*, 73(364):730–738, 1978. ISSN 01621459. doi: 10.2307/2286266. URL http://dx.doi.org/10.2307/2286266.

J. D. Ramsey, S. J. Hanson, C. Hanson, Y. O. Halchenko, R. A. Poldrack, and C. Glymour. Six problems for causal inference from fMRI. *NeuroImage*, 49(2):1545–1558, 2010. URL http://dblp.uni-trier.de/db/journals/neuroimage/neuroimage49.html#RamseyHHHPG10.

R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, 1984. ISSN 00361445. doi: 10.2307/2030064. URL http://dx.doi.org/10.2307/2030064.

T. Richardson. A discovery algorithm for directed cyclic graphs. In *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence*, UAI'96, pages 454–461, San Francisco, CA, USA, 1996. Morgan Kaufmann Publishers Inc. ISBN 1-55860-412-X. URL http://dl.acm.org/citation.cfm?id=2074284.2074338.

T. Richardson and P. Spirtes. Ancestral graph Markov models. *Annals of Statistics*, 30: 2002, 2000.

A. H. Roffo. Durch tabak beim kaninchen entwickeltes carcinom. *Zeitschrift für Krebsforschung*, 33:321, 1931.

A. H. Roffo. Krebserzeugendes benzpyren, gewonnen aus tabakteer. *Zeitschrift für Krebsforschung*, 49(5):588–597, 1939. ISSN 1432-1335. doi: 10.1007/BF01620960. URL http://dx.doi.org/10.1007/BF01620960.

E. Schairer and E. Schöniger. Lungenkrebs und tabakverbrauch. *Zeitschrift für Krebsforschung*, 54(4):261–269, 1944. ISSN 1432-1335. doi: 10.1007/BF01628727. URL http://dx.doi.org/10.1007/BF01628727.

E. Sgouritsa, D. Janzing, J. Peters, and B. Schölkopf. Identifying finite mixtures of nonparametric product distributions and causal inference of confounders. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 556–565, Oregon, USA, 2013. AUAI Press Corvallis.

S. M. Smith, K. L. Miller, G. S. Khorshidi, M. A. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, and M. W. Woolrich. Network modeling methods for fMRI. *NeuroIm-*

age, 54(2):875–891, 2011. URL http://dblp.uni-trier.de/db/journals/neuroimage/neuroimage54.html#SmithMKWBNRW11.

P. Spirtes. Conditional independence properties in directed cyclic graphical models for feedback. Technical report, Carnegie Mellon University, 1994.

P. Spirtes. Directed cyclic graphical representations of feedback models. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, pages 491–498, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1-55860-385-9. URL http://dl.acm.org/citation.cfm?id=2074158.2074214.

P. Spirtes and T. Richardson. A polynomial time algorithm for determining DAG equivalence in the presence of latent variables and selection bias. In *Proceedings of the 6th International Workshop on Artificial Intelligence and Statistics*, 1996.

P. Spirtes, C. Meek, and T. Richardson. An algorithm for causal inference in the presence of latent variables and selection bias. In *Computation, Causation, and Discovery*, pages 211–252. AAAI Press, Menlo Park, CA, 1999.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

D. Stoyan, W. Kendall, and J. Mecke. *Stochastic geometry and its applications*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley, 1987. ISBN 9780471905196. URL https://books.google.com/books?id=DU3vAAAAMAAJ.

E. V. Strobl, K. Zhang, and S. Visweswaran. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. 2017. URL http://arxiv.org/abs/1702.03877.

I. Tsamardinos, L. E. Brown, and C. F. Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, Oct. 2006. ISSN 0885-6125. doi: 10.1007/s10994-006-6889-7. URL http://dx.doi.org/10.1007/s10994-006-6889-7.

C. Uhler, G. Raskutti, P. Bühlmann, B. Yu, et al. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, 41(2):436–463, 2013.

C. F. J. Wu. On the convergence properties of the EM algorithm. *Ann. Statist.*, 11(1): 95–103, 03 1983. doi: 10.1214/aos/1176346060. URL http://dx.doi.org/10.1214/aos/1176346060.

E. L. Wynder, E. A. Graham, and A. B. Croninger. Experimental production of carcinoma with cigarette tar. *Cancer Research*, 13(12):855–864, 1953. ISSN 0008-5472. URL http://cancerres.aacrjournals.org/content/13/12/855.

I. Yeo. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959, 2000.

M. I. M. Yusoff, M. R. A. Bakar, and A. H. S. M. Nor. The performance of expectation maximization (EM) algorithm in Gaussian mixed models (GMM). *Pertanika Journal of Science & Technology*, 17(2), 2009.

J. Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, Nov. 2008. ISSN 0004-3702. doi: 10.1016/j.artint.2008.08.001. URL http://dx.doi.org/10.1016/j.artint.2008.08.001.

K. Zhang, J. Peters, D. Janzing, and B. Schlkopf. Kernel-based conditional independence test and application in causal discovery. In F. G. Cozman and A. Pfeffer, editors, *Proceedings of the Twenty Seventh Conference on Uncertainty in Artificial Intelligence*, pages 804–813. AUAI Press, 2011. ISBN 978-0-9749039-7-2. URL http://dblp.uni-trier.de/db/conf/uai/uai2011.html#ZhangPJS11.

K. Zhang, B. Huang, J. Zhang, B. Scholkopf, and C. Glymour. Discovery and visualization of nonstationary causal models. 2015. URL https://arxiv.org/abs/1509.08056.

K. Zhang, B. Huang, J. Zhang, C. Glymour, and B. Schölkopf. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *Proceedings of the Twenty Sixth International Joint Conference on Artificial Intelligence*, 2017.